MEASUREMENT AND PREDICTION

STUDIES IN SOCIAL PSYCHOLOGY
IN WORLD WAR II

Volume I. The American Soldier: Adjustment During Army Life
Volume II. The American Soldier: Combat and Its Aftermath
Volume III. Experiments on Mass Communication
Volume IV. Measurement and Prediction

The four volumes in this series were prepared and edited under the auspices of a Special Committee of the Social Science Research Council, comprising

Frederick Osborn, *Chairman*
Leonard S. Cottrell, Jr
Leland C. DeVinney
Carl I. Hovland
John M. Russell
Samuel A. Stouffer
Donald Young, Ex Officio

The data on which these volumes are based were collected by the Research Branch, Information and Education Division, War Department, during World War II. In making the data available, the War Department assumes no responsibility for the analyses and interpretations contained in these volumes, which are the sole responsibility of the authors.

These volumes were prepared under a grant from the Carnegie Corporation of New York. That corporation is not, however, the author, owner, publisher, or proprietor of this publication, and is not to be understood as approving by virtue of its grant any of the statements made or views expressed therein

VOLUME IV

MEASUREMENT AND PREDICTION

BY

SAMUEL A. STOUFFER

LOUIS GUTTMAN

EDWARD A. SUCHMAN

PAUL F. LAZARSFELD

SHIRLEY A. STAR

JOHN A. CLAUSEN

PRINCETON, NEW JERSEY

PRINCETON UNIVERSITY PRESS

1950

London: Geoffrey Cumberlege, Oxford University Press

Printed in the United States of America by
The Colonial Press Inc., Clinton, Mass.

FOREWORD

This volume organizes some new thinking and research on measurement and prediction in social psychology and sociology. It grew out of the work of the Research Branch of the Information and Education Division in the War Department in World War II.

Back of the substantive findings of the Research Branch, reported in Volumes I and II of this series, stands a wide variety of methodological research. Selected contributions are described in detail in Volume III and in the present volume.

In Volume III, methodological problems involved in the use of controlled experiments are analyzed in some detail, especially as they apply to the field of mass communication.

In Volume IV the first eleven chapters deal with a theoretical and empirical analysis of problems of measurement—particularly the development of models of ordered structures or scales and practical procedures for testing their adaptability to a given area of socio-psychological content. Chapters 12 through 16 deal with problems of prediction as illustrated by two of the major efforts of prediction made by the Research Branch.

Chapter 1 provides an introductory overview to the contributions to measurement and Chapter 12 performs the same office for the contributions to prediction.

These reports leave out other fields of methodological endeavor in which the Research Branch was interested. To cite one example, efforts were made to improve methods of interviewing. Special attention was given to informal and unstructured or partly structured interviews, as supplementary to the more standard questionnaire techniques generally used in the Branch. Some of this experience has been published elsewhere.[1]

As the introductory chapter in Volume I of this series pointed out, the mission of the Branch was one which might be called an engineering rather than a scientific one. To carry out the practical

[1] See, for example, Robert K. Merton and Patricia L. Kendall, "The Focused Interview," *American Journal of Sociology*, Vol. 51, No. 6 (May 1946), pp. 541–557.

engineering tasks, there was need for better research techniques. Work on these techniques was, of necessity, incidental to the main responsibility of getting reasonably accurate information quickly into the hands of those who needed the information for policy decisions. Hence, many of the social scientists, being keenly sensitized to the shortcomings of techniques in current use, were often frustrated at not having more time to turn away from immediate problems to intensive study of these methodological challenges. Nevertheless, some progress was made, as will be seen.

In Volume I, Chapter 1, the view was expressed that the future of social psychology and sociology calls for three developments:

1. Formulation of theories, at least of some limited generality, which can be operationally stated such that verification is possible, and from which predictions can be made successfully to new specific instances.

2. Such theories demand that the objects of study be isolated and accurately described, preferably by measurement.

3. Once the variables are identified, the test of adequacy of the theory, in comparison with alternative theories, must be rigorous, preferably evidenced by controlled experiment, and preferably replicated.

By its contributions to measurement and prediction this volume seeks to accelerate the advance of social science.

This is not a textbook. Neither is it a comprehensive treatise. Rather, it organizes, as compactly as possible, a considerable amount of fresh thinking on these problems, with copious illustrations from Research Branch data.

While the basic concepts on which this volume is based developed in large part in response to need for methodological improvement of practical research operations in wartime, thinking has not stood still since the war. This volume incorporates subsequent reflection and research, although all empirical data are from Research Branch materials. In the measurement field, for example, Louis Guttman has extended and generalized the approaches which he initiated for use in the War Department, and Paul F. Lazarsfeld, whose work began in the closing stages of the war, has developed most of his systematic theory subsequently. Completion of this volume has been made possible by a grant to the Social Science Research Council by the Carnegie Corporation and by assistance from other sources. The Harvard University Laboratory of Social Relations and the Columbia University Bureau of Applied Social Research have con-

tributed facilities and funds. In 1948–1949 the RAND Corporation under Air Force Project RAND made available to Harvard University a grant for further theoretical and empirical research on socio-psychological measurement. While only partial findings of the RAND study are available as this volume goes to the printer, the experience on this study has been of inestimable value in broadening the perspective of several of the authors of the present volume and leading them to make late revisions in their manuscripts—in some cases, revisions which are quite fundamental in character. Other contributions to this volume also draw on postwar research. John A. Clausen's Chapters 15 and 16 on the prediction of what veterans would do after discharge, for example, depend in part upon data collected by the Veterans Administration as a follow-up of the same individuals studied by the Research Branch.

Special acknowledgment is due to Frederick Mosteller, who contributed wise mathematical criticism to several of the chapters. Preparation of the final manuscript was the responsibility of Stuart Cleveland. The major proofreading task was performed by Blanchard Lyon, who also prepared the index.

While this is a methodological volume, the content material used for illustrative purposes in some of the chapters has an intrinsic substantive interest. For example, in Chapters 13 and 14, by Shirley A. Star, tables are presented on the scores made on a psychoneurotic inventory by all men in the United States inducted during a single month. Data are given separately for each induction station. These findings conceivably may have considerable impact on psychiatry, since they provide for the first time a bench mark against which to interpret the almost fantastic variations from one induction station to another throughout the nation in psychiatrists' diagnoses.

Most of the chapters can be read by the nonmathematician. But some chapters demand mathematical literacy, and all are written for the serious student.

We expect some of the topics to be rather highly controversial. Several chapters represent a challenge to conventional thinking in this field. But they are not conceived by their authors as definitive. It is the hope that the study of these pages will inspire a younger generation of social scientists to do better the tasks here undertaken.

SAMUEL A. STOUFFER

CONTENTS

MEASUREMENT AND PREDICTION

CHAPTER 1

AN OVERVIEW OF THE CONTRIBUTIONS TO SCALING AND SCALE THEORY[1]

Introduction

VITAL to the development of social psychology and sociology as scientific disciplines is the definition and classification of the objects of study.

This requires rigorous yet economical methods for handling data which are initially *qualitative*, not quantitative. The objective of much of the Research Branch methodological endeavors, and of the first eleven chapters of this volume, is to deal *with theoretical models of ordered structures or scales and with technical procedures for testing the applicability of a particular model to a particular set of qualitative data.*

In the course of this work a number of general theories have been developed which go well beyond present computational possibilities but which may have important implications for the future.

Some of these results have already been published. Others have been presented as papers before professional societies. In these chapters the work is brought together in one place, in a form which, it is hoped, will be maximally useful both to the researcher who wants to make practical use of the new tools and to the scholar who seeks to sharpen his knowledge of the shortcomings of these tools in order to develop better ones.

It is to be expected that considerable sections of these chapters will be controversial. The critical papers which have already ap-

[1] By Samuel A. Stouffer. While the author has had the benefit of critical readings of drafts of this chapter by the authors of Chapters 2 to 11, namely, Louis Guttman, Edward A. Suchman, and Paul F. Lazarsfeld, as well as by other experts in the field, he must assume responsibility for the summary statements in the present chapter. He has sought as faithfully as possible to represent the points of view of these authors and to reconcile divergencies of opinion, though not always succeeding completely, perhaps, to each author's satisfaction.

peared, based on fragmentary advance publications, are testimony to the vitality of the problems with which these chapters deal.[2]

Original and challenging as some of the new concepts here introduced may be, they represent only one small body of effort in a task whose magnitude is likely to absorb the energies of some of our best minds in psychology and sociology for many years to come. It is a task which has elicited effort from able scholars long before the war.

The thinking in these chapters owes a special debt to the prewar work of L. L. Thurstone of the University of Chicago, whose adaptation of psychophysical methods to attitude measurement was one of the earliest achievements in combining systematic theory and computational techniques in this area and whose subsequent leadership in the field of factor analysis built upon the work of Spearman and opened new vistas in psychological measurement generally. There are many other scholars whose research added to our knowledge. No systematic attempt will be made here to summarize this literature. It is assumed that the reader has some familiarity with it and, in any case, a number of summaries are available.[3]

Most of that work represented an effort to apply quantitative methods to qualitative data. While the reader will soon discover that the present volume involves some quite diverse approaches based on diverse premises, the one thing which these chapters have in common, differentiating them from the prewar work, is their effort to treat qualitative data *as* qualitative, not quantitative.

Perhaps the most drastic departure from earlier approaches is represented in the initial thinking of Louis Guttman. In 1940, just before the war, Guttman contributed a series of studies on the logic of measurement and prediction to a monograph of the Social Science Research Council.[4] This work contained the basic principle of ideas

---

[2] For a discerning summary of this literature, see Jane Loevinger, "The Technic of Homogeneous Tests Compared with Some Aspects of 'Scale Analysis' and Factor Analysis," *Psychological Bulletin*, Vol. 45, No. 6 (November 1948), pp. 507–529.

[3] Thurstone's approaches to attitude measurement based on psychophysical methods are summed up in a recent paper by him published as Chapter 5 in T. G. Andrews (editor), *Methods in Psychology* (John Wiley & Sons, Inc., New York, 1948). On his factor analysis, see L. L. Thurstone, *Multiple Factor Analysis* (University of Chicago Press, Chicago, 1947). The best prewar elementary textbook on measurement in psychology is probably J. P. Guilford, *Psychometric Methods* (McGraw-Hill Book Company, New York, 1936). David Krech and Richard S. Crutchfield, *Theory and Problems of Social Psychology* (McGraw-Hill Book Company, New York, 1948) has a relatively up-to-date bibliography in Chapter 7, "The Measurement of Beliefs and Attitudes." One of the most complete bibliographies of the literature on attitudes in general is that in Muzafer Sherif and Hadley Cantril, *The Psychology of Ego-Involvements* (John Wiley & Sons, Inc., New York, 1947).

[4] Louis Guttman, "The Quantification of a Class of Attributes: A Theory and Method for Scale Construction," in P. Horst et al., *The Prediction of Personal Adjustment* (Social Science Research Council. New York, 1941), pp. 319–348.

which he was to apply in the Research Branch a year later and which were to be greatly expanded theoretically and adapted for quick practical use.

Guttman offered a model which dispenses with the concept of a latent or underlying continuum to which the response to a particular item is to be related. He considered an attitude area "scalable" if responses to a set of items in that area arranged themselves in certain specified ways. In particular, it must be possible to order the items such that, ideally, *persons who answer a given question favorably all have higher ranks than persons who answer the same question unfavorably*. From a respondent's rank or scale score we know exactly which items he endorsed. Thus we can say that the response to any item provides a *definition* of the respondent's attitude.

Guttman and his associates in the Research Branch developed simple and practical techniques for testing hypotheses as to whether attitude areas were scalable by this definition. Not all attitude areas satisfied the criteria of goodness of fit of the model, but it was possible to find many areas which seemed to be scalable and several hundred such scales were worked out during Research Branch experience. Also important, theoretically, is the fact that the ranking of people by the Guttman model appears to represent only one of a set of *principal components*. In fact, if there are m scale types or rank groups, in the case of perfect scalability, there are m principal components which Guttman found to have a definite law of formation. The first component is a monotonic increasing function of the ranks—a straight line in the special case where the frequencies are the same for each rank group. The second component will always be a U-shaped or J-shaped curve with one bend in it. This has been interpreted by Guttman as a measure of *intensity*, leading, in the ideal case, to the determination of an objective zero point, at the point where the curve of intensity is a minimum.

The perfect Guttman scale is not likely to be found in practice, but satisfactory approximations to it are not so rare as some critics of the early papers describing it have implied. In analyzing data which failed to fit the model, Guttman has distinguished between two classes of misfits. One, which he calls the *quasi scale*, has the property that the errors are at random. He proves that the correlation of the quasi scale with an outside criterion is the same as the multiple correlation between responses to the individual items forming that scale and the outside criterion. This important result leads to great operational economies and justifies the use of sets of items from an area not scalable in his strictest sense. If the condition of

random errors is not satisfied we are confronted with what are called *nonscale types*. These nonscale types set problems for further analysis, by indicating the presence of more than one variable. Intensive study in any given situation may lead to the dissection of a supposed attitude area into two or more subareas each of which may turn out to be scalable, in Guttman's sense, when analyzed separately.

Further research is needed on the quasi scale viewed simply as a system of manifest responses, without necessarily postulating an explicit underlying or latent continuum.

In the last years of the war, Paul F. Lazarsfeld, while a consultant in the Research Branch, became especially concerned with these quasi scales. For several years he had been interested in the classification of attributes, and especially in the properties of partial fourfold tables, on which Yule had written many years before with much insight in the opening chapters of his *Introduction to the Theory of Statistics*. As a result of his studies, Lazarsfeld proposed a fresh attack on the fundamental conceptualization of the scaling problem.

What Lazarsfeld proposed was a return to the older concept of a *latent attitude continuum*. Guttman, before the war, had studied, perhaps more thoroughly than any other person, the possibilities of this approach but had turned for the time being in a new and different direction because he saw clearly that conventional methods of quantitative factor analysis, stemming from Spearman, were inappropriate for handling qualitative data. Lazarsfeld's primary achievement was to bring forth a new model for directly factoring qualitative data. Therefore, it would now seem to be practical to test quite rigorously the following basic hypothesis: *There exists a set of latent classes, such that the manifest relationship between any two or more items on a test can be accounted for by the existence of these basic classes and by these alone.*

This implies that any attitude item has two aspects—one which is associated with the latent classes and one which is specific to the item. The specific aspect of any item is assumed to be independent of the latent classes and also independent of the specific aspect of any other item.

The analogy with the Spearman-Thurstone approach in quantitative factor analysis and indebtedness to it is obvious. The contrast with Guttman's approach also should be clear. Guttman's model deals only with the manifest relationship among attitude

items and *defines* an attitude directly as the observed responses to these items. Lazarsfeld defines an attitude as an *inference* as to latent classes, tested by fitting to the manifest data an appropriate latent structure model.

Lazarsfeld's approach is quite general as to the number of latent classes. Moreover, these classes may or may not be *ordered.* In the very important special case where they are ordered along a single dimension, Lazarsfeld has achieved further generalizations of importance. He conceives of the latent classes as segments of a continuum x. Over this latent attitude continuum, the probability, $p(x)$, of a "favorable" response to a given specific attitude item, may be described as a continuous function of x. Such a "trace line" can take any shape, but the study of relatively simple types of "trace lines," such as the straight line and polynomials of second and third degree, is proving to be of much interest.

It turns out, further, that Guttman's quasi scale can be derived analytically as a special case of latent structure analysis, Guttman's perfect scale becoming a limiting case of the quasi scale.

In comparing the approaches of Guttman and Lazarsfeld to the problem of testing whether or not ordered structures apply in a given attitude region, it is important to avoid metaphysical faith in a particular model. For example, if one can dispense with the concept of a latent structure and operate effectively directly with manifest data, as Guttman does, there could be advantages of logical parsimony in the Guttman approach. The fact that a perfect Guttman scale is not obtained except approximately is not in itself a denial of the value of that model. The crucial question is whether it has ideal properties from which one may make rich and varied logical deductions—as seems to be possible in connection with the theory of principal components—and whether it leads to rapid and enlightening empirical tests of hypotheses as to the structure of concrete attitude areas. Similarly, it will be noted that Lazarsfeld's postulate of a continuous trace line $p(x)$ cannot itself be empirically demonstrated. However, as will be shown in this volume with numerical examples from Research Branch data, the hypothetical latent continuum x can be cut into segments and discontinuous values of $p(x)$ determined empirically for each segment. The latent continuum is a hypothetical construct and its usefulness lies in its range of logical productivity and its effectiveness in guiding the analysis of actual data.

In this volume, we have not been content merely to present the

mathematics of these new approaches to the problem of testing hypotheses of ordered structures. We have sought also to provide the details of computing techniques by which various hypotheses can be tested empirically. This volume contains many concrete examples from Research Branch data. The development of computing equations and of short-cut methods for clerical operations is second only to the formulation of the basic theory in its importance for progress in this field.

The initial objective in applying one or another of the various proposed models is to discover whether, with a given set of items, an area is scalable. A second objective, if a scalable dimension is found, is to *order* respondents along that dimension in a practical operation.

There are many methods in current use for ordering respondents. The term "item analysis" is commonly used to describe most such operations. Usually, a provisional total score is computed by adding up a person's responses to items initially weighted, and decision as to whether or not to retain a particular item in the "scale" turns on how highly the responses to it are correlated with the provisional total score. The objection to this procedure is that, at the end, one is still likely to be retaining several dimensions in a single so-called "scale." Such a "scale" may appear to order respondents, but it is not backed by a theoretical model which provides a serviceable criterion that these respondents have been ranked along one dimension only.

The Research Branch, like most practical research agencies, faced a situation in which it was usually not economical to use more than a few items in a given attitude scale. Or, if the presence of several dimensions was suspected, it was not possible to use large batteries of tests to isolate these dimensions. Moreover, it was adequate *to order people in a few rank groups,* provided there was a single demonstrable dimension. It was not necessary, nor is it likely to be necessary in much of the social and psychological research of the near future, to have a very large number of rank groups.

One of the fortunate results of the developments reported in the following chapters is the evidence that—if an attitude area is scalable—a relatively small number of items is sufficient for classifying people for many practical purposes. Thus, while a larger number of items may be required in pretests in order to study the structure of each attitude, only a small number of items need be used in a final study for those attitudes which are scalable.

It may be objected that the price one pays for unidimensionality is loss of generality. That can be true. But, if a given region of attitudes is not unidimensional it does no good to close our eyes to that fact and pretend that it is. Instead, if the region yields several quite specific scales, each clearly unidimensional, we are then in a position to proceed more economically with an explicitly multiple dimensional study, testing theories of the interrelationships of the scales used. This study may require us to try eventually to reduce the number of these initial dimensions by some form of quantitative factor analysis. But the first step is analysis of the structure, in an effort to isolate clean unidimensional specific scales.

In order to give the reader further orientation as to the viewpoints and findings of the chapters to follow, we shall now review in somewhat more detail, without mathematics, (1) the Guttman approach to the testing of hypotheses as to ordered structures and (2) the latent structure approach as developed by Lazarsfeld. These will be viewed in terms of the objective of finding among respondents a rank order along a particular attitude continuum, if such exists. Finally, (3) we shall take a preliminary look at the problem of utilizing indices of intensity of attitude to locate a "zero point" or "region of indifference."

1. *The Scalogram Theory of Establishing Rank Order*

The approach which was developed in the Research Branch under the guiding hand of Louis Guttman has been named scalogram analysis. Several attitude areas were analyzed during the war to see whether items from these areas could be accepted as scalable. Some of them have been used in earlier volumes of this series. As experience proceeded, some of the earlier criteria were revised, and many of the areas once deemed sufficiently scalable do not satisfy the more rigorous criteria subsequently imposed. But the general logical outline, which is quite a simple one, has stood the test of a wide variety of applications.

The scalogram hypothesis is that the items have an order such that, ideally, *persons who answer a given question favorably all have higher ranks on the scale than persons who answer the same question unfavorably*. From a respondent's rank or scale score we know exactly which items he endorsed. Thus, ideally, scales derived from scalogram analysis have the property that the responses to the individual items are reproducible from the scale scores.

In practice, this ideal is not perfectly attained. In Chapters 3

and 4 standards are set forth for accepting data as constituting such a scale. There have been some misunderstandings in the profession as to these standards, and as to the reasoning as well as the empirical experience on which they are based.

The standards have been and doubtless will be criticized from two quite opposite points of view: (a) that they are too lenient, permitting areas to be accepted as scalable when they are not, and (b) that they are too stringent, limiting too severely the number of areas which can be accepted as scalable. There is an arbitrary element of judgment involved in arriving at a standard, just as there is in arriving at a standard such as the 5 per cent level for a test of significance in sampling theory. It is quite possible that some of the standards here proposed may be changed with increasing experience. Research Branch experience has shown that when the proposed criteria are adopted, many areas have been found to be scalable; rank scores based on even a small number of items from such an area necessarily have high test-retest reliability (according to a theorem of scalogram analysis) and have been found in practice to correlate satisfactorily with external variables.

The items used in a scalogram analysis must have a special *cumulative property*. The general idea may be illustrated by a hypothetical scale of stature comprised of responses to three items:

1. Are you over 6 feet tall?	______ Yes	______ No
2. Are you over 5 feet 6 inches tall?	______ Yes	______ No
3. Are you over 5 feet tall?	______ Yes	______ No

If a person checks item 1 "Yes," he must, unless he is careless, also check items 2 and 3 "Yes." If he checks item 1 "No" and item 2 "Yes" he must also check item 3 "Yes." Hence, if we give a score of 2 to a man who has endorsed two items we know exactly *which two items* he endorsed. He could not say "Yes" to item 1, "No" to item 2, and "Yes" to item 3. The four admissible response patterns to the three items are shown below:

Rank order of respondents	*Score*	*Says yes to item*			*Says no to item*		
		1	*2*	*3*	*1*	*2*	*3*
1	3	X	X	X			
2	2		X	X	X		
3	1			X	X	X	
4	0				X	X	X

This simple diagram is called a scalogram—hence the name scalogram analysis for the procedure.

It is essential to note that the items are *cumulative.* A man who answers "Yes" to item 1 has the stature of a man who answers "Yes" to item 2 and "No" to item 1, *plus additional stature.* Thus the following items would not yield a scale in a scalogram analysis. They are not cumulative.

1. Are you over six feet tall?	______ Yes	______ No
2. Are you between 5 feet 6″ and 6 feet tall?	______ Yes	______ No
3. Are you between 5 feet and 5 feet 6″ tall?	______ Yes	______ No
4. Are you under 5 feet tall?	______ Yes	______ No

Here it is possible to make only one affirmative response and we must use additional knowledge of the fact that a man who says "Yes" to item 1 is taller than a man who says "Yes" to item 2 if we are to order the items. The scale picture would look as follows:

Rank order of respondents	*Says yes to item*				*Says no to item*			
	1	*2*	*3*	*4*	*1*	*2*	*3*	*4*
1	X					X	X	X
2		X			X		X	X
3			X		X	X		X
4				X	X	X	X	

In this example, the order of the items is implicit in the structure of the questions. Where this is not the case, a device sometimes used is to have a set of judges rank the items in advance. The consistency with which the judges agree in their initial ranking can be evaluated. But since only one answer is permissible, there is no test of internal consistency of the subsequent *respondents.* Such a test could be devised by, for example, asking each respondent to check "Yes" to the *two* items closest to his position. This would reduce the number of rank groups, but would provide objective evidence of order such that initial judges would no longer be needed, although they might still be useful. The scale picture would be:

Rank order of respondents	*Says yes to item*				*Says no to item*			
	1	*2*	*3*	*4*	*1*	*2*	*3*	*4*
1	X	X					X	X
2		X	X		X			X
3			X	X	X	X		

It would seem to be quite possible to use as a model this type of "interval" or "position" scale. In fact, Frederick Mosteller has developed the theory of such a scale, in connection with a study of scaling sponsored by the RAND Corporation at the Harvard Laboratory of Social Relations. It will be noted that the cumulative type of item which fits the scalogram model will not fit this model, and vice versa.

One can handle the stature items 1, 2, 3, and 4 above in still other ways. For example, the respondent might be asked to rank the four stature intervals in terms of their closeness to his height. If he is 5 feet, 2 inches tall, the order would be interval 3, rank 1; interval 4, rank 2; interval 2, rank 3; interval 1, rank 4. For all respondents, we would get the following scale picture (the man whose stature is 5 feet, 2 inches, is in rank group 5, for example):

Rank order of respondents	*Rank assigned to item number*			
	1	*2*	*3*	*4*
1	1	2	3	4
2	2	1	3	4
3	3	1	2	4
4	4	2	1	3
5	4	3	1	2
6	4	3	2	1

The same result could be derived from a set of $n(n - 1)/2$ paired comparisons, which would have the additional advantage of providing a check on the internal consistency of an *individual* as well as the internal consistency of a *group* of *individuals*. Questions would be of the type:

Which of these is closer to your stature? (Check one)
______ Over 6 feet
______ Between 5 feet 6″ and 6 feet

All of these examples illustrate how a model can be set up without *explicitly* postulating a *latent* continuum. The ordering of manifest responses *if perfectly consistent* will reveal the appropriate structure.[5]

[5] Thurstone's law of comparative judgment postulates an underlying continuum and seeks to derive, usually with the aid of the method of paired comparisons, not only a rank order but a metric. In psychophysical measurement, such as the classical example of lifted weights, the assumption is made that each individual respondent has the same position and that variations in judgment either within the same individual or between a group of individuals represent discriminal error. This discriminal error provides a unit of measurement. But in the present example, each individual can be perfectly consistent in all his responses, yet because each individual can have a different position

The scalogram model, requiring cumulative items, has at the present time been studied far more thoroughly than these others. Therefore, our attention will now be focused on it. We must keep clearly in mind the fact that the cumulative character of the items satisfying a scalogram analysis puts a restriction upon the type of data which can be used with this model. Yet the restriction is not, in practice, as severe as may at first appear.

It is often possible to find items which have an intrinsic cumulative character. The prototype is perhaps the social distance scale, with such items as the following:

1. Would you want a relative of yours to marry a Negro?
2. Would you invite a Negro to dinner at your home?
3. Would you allow a Negro to vote?

An illustration from Research Branch data of this type of intrinsic scale may be taken from a study of riflemen overseas who had recently experienced combat. The data are presented more in detail in Chapter 5. It was desired to see if there was a scale order in the way in which respondents reported experiencing fear. The following question was asked:

> Soldiers who have been under fire report different *physical reactions to the dangers of battle.* Some of these are given in the following list. How often have you had these reactions when you were under fire? Check one answer after each of the reactions listed to show how often you had the reaction. Please do it carefully.
>
> [There followed 10 items, each with a four-step check list. For example:
>
> Shaking or trembling all over
> ______ Often
> ______ Sometimes
> ______ Once
> ______ Never]

A scale picture (shown in Chapter 5, Scalogram 7) indicated that nine of these items formed a very satisfactory scale when ordered as follows:

1. Urinating in pants
2. Losing control of the bowels

(i.e., different stature), the group variability will not represent *discriminal* error. Detailed analysis of this phenomenon has been made by Clyde H. Coombs, in connection with his work on the RAND study at the Harvard Laboratory of Social Relations. Among other things, he has shown that a rank order table like that above, or a paired comparisons table, does reveal under certain conditions whether some rank intervals have greater magnitude than others.

3. Vomiting
4. Feeling of weakness or feeling faint
5. Feeling of stiffness
6. Feeling sick at the stomach
7. Shaking or trembling all over
8. Sinking feeling of the stomach
9. Violent pounding of the heart

One item—"cold sweat"—did not fit into the scale; that is, some people who experienced less frequent fear symptoms than cold sweat also experienced cold sweat, but this was also true of some people who experienced more frequent symptoms than cold sweat. Hence "cold sweat" was shown to involve a factor or factors additional to the scale variable.

The fact that the nine items satisfied the scalogram criteria means that if a man did not report vomiting, for example, he also did not report urinating in his pants or losing control of his bowels. If he did report vomiting, he also reported all of the other experiences (4 through 9 on the list) which were generally more frequent than vomiting. In this case he would have a score of 7 and if the scale were perfect we would know exactly *which* seven experiences he reported. Actually, of course, the scale was not quite perfect. By an easy procedure, described in Chapter 4, it was possible to compute a *coefficient of reproducibility*, which was .92. This means that if we knew any scale score, such as 7, and if we guessed the exact items which respondents with this scale score endorsed and did not endorse, we would guess 92 out of 100 of the items correctly and 8 out of 100 incorrectly. The scale picture also met other criteria described in Chapters 3 and 4, with respect to individual item reproducibility, wide range of marginals, and randomness of error.

It may be that only a limited range of psychological and social phenomena have the intrinsic cumulative characteristic described above. An area which should eventually be quite rich in data of this type should be that of level of difficulty. For example, three problems in calculus:

1. Integrate $\frac{dy}{dx} = xy(y - a)$
2. Integrate $\frac{dy}{dx} = x^2$
3. Differentiate $y = x^2$

It is unlikely that a person who can do problem 1 would fail on problems 2 and 3, and it is also unlikely that a person who failed on prob-

lem 2 could do problem 1. In other words, if A has a higher score than B, A can do all that B can do, *plus* something more.

The possibility should be faced, however, that large areas of psychological or social behavior may not yield items which approximate the scale pattern required by scalogram analysis. For example, a short catalogue of psychosomatic. symptoms which the Research Branch used in constructing a psychiatric screening test simply did not order itself in a manner satisfying the rigorous criteria set up. That is, a man who said he was bothered with shortness of breath might or might not complain of spells of dizziness, and vice versa. As we shall see later, it is still possible to construct a scale—called a *quasi scale*—which lacks the property of reproducibility but which does have the valuable property of *yielding the same correlation with an outside criterion as does the multiple correlation of the individual items with that criterion.*

It is frequently possible, however, to order items cumulatively (and hence to order respondents) by constructing items with multiple check lists and choosing cutting points by combining categories such that the error of reproducibility is minimized.

Consider, for example, three items which happen to have uniform format. (Such uniformity is not at all essential—the number of response categories can vary arbitrarily from item to item and the wording of the categories can be varied):

1. How many of your officers take a personal interest in their men?
   1 ______ All of them
   2 ______ Most of them
   3 ______ About half of them
   4 ______ Some of them
   5 ______ Few or none of them

2. How many of your officers will go through anything they ask their men to go through?
   1 ______ All of them
   2 ______ Most of them
   3 ______ About half of them
   4 ______ Some of them
   5 ______ Few or none of them

3. How many of your officers are the kind you would want to serve under in combat?
   1 ______ All of them
   2 ______ Most of them
   3 ______ About half of them
   4 ______ Some of them
   5 ______ Few or none of them

If each item is dichotomized by taking the two top categories as "favorable" responses, the three items will not ordinarily form a scale by scalogram analysis. In a sense, in spite of manifest difference in content, these three particular items are more or less synonymous. Hence, the frequencies probably will not be cumulative. However, if for item 1 we take response category 1 as "positive," for item 2 response categories 1 and 2, and for item 3 response categories 1, 2, and 3, we perhaps can build a cumulative set of items. This is, of course, just one of various ways in which a cumulative set of responses might be built from these items. The procedure for selecting cutting points which will maximize scalability has been reduced to a simple routine and is described in detail in Chapter 4. It is *not arbitrary*, but is based upon a study of the *simultaneous distribution* of all the original responses.[6] From a practical standpoint, the operational procedures developed in the Research Branch for swiftly evaluating a large number of questions simultaneously may rank as one of the major contributions of the Branch to social science technique. Work which would have required hundreds of hours of elaborate machine analysis can now be done in a few hours by one semi-skilled clerk. The reader will find a full description of the operations in Chapters 4 and 5, with a diagram for the construction of a scalogram board, which remarkably facilitates the scaling process.

There can be no doubt that the empirical choice of cutting points on more or less synonymous items in order to achieve a set of cumulative responses is rather crude. It needs to be studied and criticized. But it does supply an ordering of respondents with a high degree of reliability and the ultimate test will be its utility in comparison with other techniques.

Relatively few of the scales described in this volume can be strictly described as based on near synonymous items of the kind illustrated above. In Chapter 4, for example, the detailed procedures used in working out a scale of general attitudes toward the Army are described. Twelve items comprise this scale, and some of the items

[6] Two people can check different response categories and still have the same attitude, because they differ in verbal habits. For example, one person may have a general tendency toward extreme statements and say "all of them" to a given question, whereas another person may have the same attitude toward officers but express it on the same question by answering "most of them." If these two responses are treated as separate categories they could exhibit substantial error of reproducibility, which tends to vanish when the categories are combined (that is, treated as though they were the same response). The scalogram analysis shows how best to combine categories in order to reduce errors of reproducibility.

rather closely resemble each other in format. But they are certainly not synonymous. Consider four of them:

In general, do you think you yourself have gotten a square deal in the Army?
______ Yes, in most ways I have
______ In some ways yes, in other ways, no
______ No, on the whole I haven't gotten a square deal

In general, how well do you think the Army is run?
______ It is run very well
______ It is run pretty well
______ It is not run so well
______ It is run very poorly
______ Undecided

Do you think the Army has tried its best to look out for the welfare of enlisted men?
______ Yes, it has tried its best
______ It has tried some, but not enough
______ It has hardly tried at all

Do you think when you are discharged you will go back to civilian life with a favorable or unfavorable attitude toward the Army?
______ Very favorable
______ Fairly favorable
______ About 50–50
______ Fairly unfavorable
______ Very unfavorable

These items, along with eight other items also of a general character but varying in format, were found to cumulate, after the choice of appropriate cutting points, such that they satisfied the scalogram criteria. The operations, step by step, are described in Chapter 4.

The greater the variety of questions, of course, the wider the generalization which can be made about the coverage of a particular scale, assuming that criteria of scalability are satisfied.

It is recommended in Chapters 3 and 4 that a relatively large number of items, preferably as many as ten or twelve, be used in the initial testing of the hypothesis of scalability, and that, if possible, some of these items be trichotomized rather than dichotomized. This will ordinarily make it quite difficult to achieve high reproducibility, but protects against spurious results which might be obtained by chance with only three or four dichotomous items.

Critics of the use of only three or four items seem to have overlooked an important distinction, namely, that while it is desirable to use as many as ten or a dozen items in initial testing of scalability, a smaller number of items can be safely selected from the scalable list for practical research purposes. If a dozen items re-

veal a scale pattern, then a smaller number selected from these dozen will show the same pattern (though with fewer ranks). Because they belong to the same scale, according to scalogram criteria, a smaller subset of items finally selected for use—possibly comprising only three or four—will of necessity correlate very highly with other subsets of items in the same scale, as is proved in Chapter 8 and illustrated with numerical examples. Items comprising such a subset will have higher reliability than similar items from most other types of scales, since each response is a *definition* of the respondent's position on a single continuum and is minimally corrupted with other affective material which correlates with some items and not with others.

In fact, one of the important consequences of finding an attitude to be scalable is that one is then justified in selecting three or four items which can be used to order respondents in a limited number of ranks. It may be possible eventually to use a pretest for selecting *a single item* for practical use—such as is the staple of much conventional market research or public opinion polling. But now the item is selected *with full knowledge* as to its place in the attitude structure. Intensity analysis in particular, as discussed in the last section of this chapter, may help determine objectively which item is unbiased with respect to the attitude as a whole, in the sense of dividing people into two groups, those favorable and those unfavorable on a given issue.

In attitude research of the future, an important desideratum will be to obtain simultaneous measures on *a large number of continua* from a given respondent. Scalable attitudes lend themselves particularly well to this in practice. Why? Because each attitude can be represented by only a few items, so that a large complex of attitudes can be observed in a single study. Thus multiple factor analysis might be employed to construct a typology of these continua, and we may well be on the road to an analysis of socio-psychological problems comparable to the road which has led to progress in the study of human abilities.

When the hypothesis of the scale structure is found not to be borne out by the observed data, the attitude items may still have an ordered structure, as has been said, namely, that of the *quasi scale.* The conditions for existence of a quasi scale need more study than is provided in this volume. Here we are focusing attention on the Guttman procedure. There is, of course, the possibility that entirely different models might have fitted the data.

The examination of the properties of quasi scales in particular and of scalogram analysis in general, which was begun by Paul F. Lazarsfeld while a consultant in the Research Branch, led to the development of a fresh approach to the problem of ordering respondents, which will be described in the following section.

2. *The Latent Structure Theory*

The conceptual models which the latent structure approach provides are described in algebraic and numerical detail in Chapters 10 and 11. In the present context we shall merely sketch the outlines of the ideas, indicate something of their possible implications for the future, and provide a numerical illustration from Research Branch data. While the development of the latent structure theory grew directly out of Lazarsfeld's work as consultant in the Research Branch, and has benefited from consultation with members of the Branch, its detailed working out did not come until after the end of the war.

The latent structure approach is a generalization of Spearman-Thurstone factor analysis. The basic postulate is that there exists a set of *latent classes, such that the manifest relationship between any two or more items on a questionnaire can be accounted for by the existence of these latent classes and by these alone.* This implies that any item has two components—one which is associated with latent classes and one which is specific to the item. The specific component of any item is assumed to be independent of the latent classes and *also independent of the specific component of any other item.*

The theory is quite general as to the number of latent classes. In order to familiarize the reader with the fundamental idea, let us first consider the special case where the number of latent classes is two.

This special case, which is called *latent dichotomy* analysis, to distinguish it from the more general concept of latent structure analysis, is quite simple to understand and, as will be shown, has some interesting properties.

In applying the latent dichotomy model we seek to partition the sample into two classes of respondents—those who possess the latent character and those who do not. The latent dichotomy is, of course, an inference. Whether or not a given set of items can be fitted by a latent dichotomy has to be determined from the data, which ideally must satisfy a very rigorous criterion.

This criterion requires, as already indicated, that *all* of the relationship between any two questionnaire items be accounted for by the hypothetical latent dichotomy. Suppose that all respondents had labels which permitted us to sort them into two classes—those possessing the latent character and those not possessing it. Our criterion requires that, among those possessing the latent character, there be no association or correlation between the responses to any two individual items. Similarly, among those not possessing the latent character, there be no association or correlation between the response to any two individual items. We shall make this concrete with a numerical example presently.

Actually, as we shall see, the ideal criterion conditions for a latent dichotomy are not likely to be realized. But they may be *approximately* realized—at least with a close enough approximation to satisfy us that the model is appropriate for a given set of items. If so, we can then use the information obtained to order the respondents.

For each response pattern—for example, $+ - + -$, a pattern in which a person is favorable on item 1, unfavorable on item 2, favorable on item 3, and unfavorable on item 4—we can estimate how many possess the latent character and how many do not We can therefore calculate an estimate of the *proportion* who possess the latent character. We can do this for all possible response patterns. We can then put the response patterns in rank order according to the rank order of the proportions estimated as possessing the latent character. Since the individual respondents must fall into response patterns, we have ordered the individual respondents, except that all those with the same response pattern are of course tied in rank.

It will be seen at once that the latent dichotomy concept hypothesizes a fundamentally different structure from that of the cumulative scale model.

The scalogram model considers the response to each item as providing a *definition* of the respondent's attitude on the subject of the particular attitude scale. The latent dichotomy model does not require us to call a response to a particular item a definition of the respondent's attitude. Instead, it requires that we conceive of his response as having two components—one, which is a manifestation of a latent character and the other, which is specific to the item. Thus two items endorsed by a person, such as, "The British are brave fighters" and "The British are unselfish allies" may not serve

as definitions of the *same* attitude, by scalogram analysis, but may prove to possess a common latent dichotomy, as well as residual content which is *specific* for each item, respectively. But note—the residual specific content of the two items must be unrelated. That is, among all men possessing the latent character, those who say the British are brave must be no more likely to say the British are unselfish allies than do those who do not say the British are brave. All the association between these two items must be attributable to the common latent character which they possess.

Now let us look at some numerical data, from a Research Branch attitude survey, which will help the reader see concretely what goes on when the latent dichotomy model is approximately realized.

We shall take the same four items on general attitudes toward the Army which were cited in the previous section of this chapter, using data from a sample of 1,000 noncoms, studied in October 1945. Responses are dichotomized, not as in scalogram procedure by choosing cutting points in such a way as to minimize scalogram error, but according to what seems, a priori, to be the manifest content of the response. We have:

1. Do you think the Army has tried its best to look out for the welfare of enlisted men?
   Proportion saying, "Yes, it has tried its best" .254
2. In general, do you think you yourself have gotten a square deal in the Army?
   Proportion saying, "Yes, in most ways I have" .300
3. Do you think when you are discharged you will go back to civilian life with a favorable or unfavorable attitude toward the Army?
   Proportion saying, "Very favorable" or "Fairly favorable" .374
4. In general, how well do you think the Army is run?
   Proportion saying, "It is run very well" or "It is run pretty well" .641

Now, after responses to each of these four items are dichotomized, there will be $2^4 = 16$ possible response patterns, determined from cross tabulation. In Table 1, column 1, the number of cases with each of the 16 response patterns is shown. For example, there are 75 cases with the response pattern $+ + + +$, favorable on all four items. There are 55 cases with the response pattern $+ - + +$, favorable on items 1, 3, and 4 and unfavorable on item 2, etc.

TABLE 1

SOLDIERS WITH VARIOUS RESPONSE PATTERNS RANKED ACCORDING TO PROPORTIONS POSSESSING LATENT CHARACTER OF FAVORABLE ATTITUDE TOWARD THE ARMY

Response pattern Item 1 2 3 4	Observed number of cases (1)	ADJUSTED NUMBER OF CASES: Total (2) =	Possessing latent character (3) +	Not possessing latent character (4)	Proportion possessing latent character $(5) = \frac{(3)}{(2)}$	Rank (6)
+ + + +	**75**	**72.2**	**71.9**	**0.3**	**.996**	**1**
+ + – +	42	44.0	42.7	1.3	.971	2
+ – + +	55	56.8	54.2	2.6	953	3
– + + +	**69**	**72.8**	**68.4**	**4.4**	**.940**	**4**
+ + + –	3	6.3	5.9	0.4	.938	5
+ – – +	45	43.7	32.2	11.5	.736	6
– + – +	60	59.7	40.6	19.1	.680	7
+ + – –	10	5.2	3.5	1.7	.674	8
– – + +	**96**	**90.8**	**51.6**	**39.2**	**.568**	**9**
+ – + –	8	8.0	4.5	3.5	.561	10
– + + –	16	11.4	5.6	5.8	.494	11
– – – +	**199**	**201.0**	**30.6**	**170.4**	**.152**	**12**
+ – – –	16	17.8	2.6	15.2	.148	13
– + – –	25	28.4	3.3	25.1	.117	14
– – + –	52	55.7	4.2	51.5	.076	15
– – – –	**229**	**226.2**	**2.5**	**223.7**	**.011**	**16**
Total	1,000	1,000.0	424.3	575.7		

The reader will recognize that only 5 of these 16 response patterns would constitute perfect scale types in a scalogram analysis free of error. These, shown in boldface type in Table 1, are:

+ + + +	75 cases
– + + +	69 cases
– – + +	96 cases
– – – +	199 cases
– – – –	229 cases
Total	668 cases

The other 332 cases are distributed among the 11 response patterns which would involve error by the scalogram model. According to the latent dichotomy model, however, these 11 response patterns *do not necessarily represent error.*[7]

[7] It must be kept in mind that the four items as here used are dichotomized quite differently from the way they were dichotomized by the scalogram analysis in which they served as items belonging to a cumulative scale.

The problem is to determine whether or not the respondents can be partitioned into two classes, those possessing and those not possessing the latent character. We will skip now the technical details of how that determination is made and concern ourselves only with its results. Since the latent dichotomy model ordinarily will not fit quite perfectly, one can compute for each response pattern an *adjusted frequency* which will differ somewhat from the *observed frequency*. The adjusted frequencies are of such a nature that the division into latent groups can be performed with precision. The adjusted frequencies are shown in column 2. A test of the adequacy of the latent dichotomy model is the agreement between the observed frequencies in column 1 and the adjusted frequencies in column 2. In the present example, it will be seen that the agreement is rather close. For example, in response pattern + + + + the observed number is 75, the adjusted number 72.2, an error of 2.8. The sum of such errors, irrespective of sign, for the 16 response patterns is 43.6 out of a total of 1,000 cases, or less than 5 per cent.[8]

Now let us focus our attention on the *adjusted* frequencies for the various response patterns. By technical operations whose description we shall for the present postpone, we can partition the adjusted frequencies into two components, those possessing and those not possessing the latent character. For example, take the response pattern − + + +. The 72.8 adjusted cases are partitioned into two latent classes, 68.4 who possess the latent character and 4.4 who do not. The numbers estimated as having the latent character are shown in column 3 of Table 1 and the numbers without it are shown in column 4. On any line the sum of the numbers in columns 3 and 4 add to the number in column 2.

Next, we can estimate the *proportion* in each response pattern who possess the latent character. This is shown in column 5. For response pattern − + + +, for example, this proportion is $68.4 \div 72.8 = .940$. Of course, we cannot pick out which *particular individual* possesses or does not possess the latent character.

With our estimate for each response pattern of the proportions possessing the latent character we can now estimate the *rank* of the response pattern as is shown in column 6. Our task is completed, for we now have the respondents *ordered* in 16 rank groups according to the proportion possessing the latent character.

Now let us look again at Table 1 and see empirically what is

[8] Appropriate tests of goodness of fit and acceptance standards are still under investigation.

meant by saying that there is no association between the items except that due to the latent dichotomy. Consider, for example, items 1 and 2 only. For those possessing the latent character, let us pool the results for all response patterns in which items 1 and 2 are both positive. We have, from column 3 in Table 1:

+ + + +	71.9
+ + + −	5.9
+ + − +	42.7
+ + − −	3.5
Total	124.0

Similarly, pooling the results for all response patterns in which item 1 is positive and item 2 is negative, we have:

+ − + +	54.2
+ − + −	4.5
+ − − +	32.2
+ − − −	2.6
Total	93.5

For all response patterns in which item 1 is negative and item 2 is positive we have 117.9 and for all response patterns in which both item 1 and item 2 are negative we have 88.9. Now let us form the following fourfold table:

		Item 1 −	*Item 1* +
Item 2	+	117.9	124.0
	−	88.9	93.5

Within the limits of error introduced by rounding, it will be seen that there is no association. In other words, the ratios of the numbers in the first row to the numbers in the second row are nearly identical. We get 1.3262 and 1.3262, identical to the fifth significant figure. Thus, among respondents possessing the latent character, there is no association between the specific components of items 1 and 2. Analogously, we can find that the same holds among respondents not possess ng the latent character and it holds for any pair of items, not merely items 1 and 2.

It must be noted again that the absence of association between

the specific components of any two items is achieved by basing the calculations on an adjusted number of frequencies (column 2 in Table 1). It would no longer quite hold if the percentages in column 5 were applied to the original observed frequencies in column 1. That is because the original data do not perfectly satisfy the criterion for the existence of a latent attribute. But we are making no *serious* error if we assume, say, that all the 75 cases in response pattern (+ + + +) have the same proportion possessing the latent attribute (.996) as the 72.2 estimated cases after adjustment. For, as we have seen, over the whole table less than 5 per cent of the 1,000 men would be misclassified thereby.

The reader should study Table 1 and the discussion of it carefully in order to grasp the basic logic of the latent structure theory as applied in this simplest of models, namely that of a latent dichotomy. Some words of caution are now in order. In particular, the reader may ask, what sort of confidence can one place in column 5 in Table 1—the proportions possessing the latent character among persons with a given manifest response pattern? It is necessary to issue a warning here that these proportions may be subject to several different sources of instability or sampling variability; for example:

(a) The kind of variability involved in test-retest, repeating the same four items on the same population of subjects.

(b) The kind of variability involved in using the same four items on two different populations of subjects.

(c) The kind of variability involved in substitution of a new "parallel form" of item for each item initially used.

As the number of questions satisfying the model increases, the number of scale patterns increases and the proportions for a scale pattern converge toward zero or unity. The number of persons with scale patterns such that one cannot be pretty sure that they do or do not possess the latent character tends to become fewer and fewer.

It follows from these considerations that the ordering of persons into 16 rank groups as in Table 1 is an empirical result which is subject to several kinds of unreliability. It is possible that the 16 rank orders thus derived would be considerably more stable than the rank orders derived from, say, four items which do not satisfy the latent structure criteria. However, a rigorous analytical job on this remains to be done.

The *crucial* thing accomplished by the operations exhibited in Table 1 is that it provides evidence of the degree to which the mani-

fest items satisfy criteria of *unidimensionality.* The need for such criteria we must keep at the forefront of our thinking. This is accomplished not perfectly in practice, but quite adequately with the latent dichotomy model in Table 1 or with more complicated models to be described presently.

Even if the values, in Table 1, column 5, of the proportions for a given manifest scale pattern who possess the latent character should "hop around" somewhat, under the conditions itemized above, it still seems useful to apply them in a given empirical investigation as an ordering scheme. In practice, we may wish to combine sixteen such patterns into two or three broad groups. Thus, if we inspect Table 1, we will see that the following manifest response patterns have proportions higher than .90 possessing the latent character:

Response pattern	p
+ + + +	.996
+ + − +	.971
+ − + +	.953
− + + +	.940
+ + + −	.938

From column 1 of Table 1 it will be seen that these five response patterns include 244 of the 1,000 individuals. Moreover, it will be seen that 521 of the 1,000 individuals have one of the following response patterns with values of p lower than .20:

Response pattern	p
− − − +	.152
+ − − −	.148
− + − −	.117
− − + −	.076
− − − −	.011

Now if we assume, provisionally, that the 244 with $p > .90$ are quite likely to have the latent character—call them group A—and the 521 cases with $p < .20$ are quite likely not to have the latent character—call them group C—we have two fairly pure extreme groups. The remaining 235 cases are indeterminate—call them group B.[9]

[9] It happens that group A contains all who were positive on at least three items, while group C contains all who were positive on only one item or less Thus the grouping is the same as might have been arrived at by conventional item analysis. But we now know much more about the properties of the items than ordinary item analysis would have told us. Moreover, it is quite possible that a tabulation of the kind of Table 1 would show that *a pattern like* − + + + *may actually have a smaller proportion possessing the latent character than a scale pattern like* + − − +.

For practical use, either for the purpose of extracting the top or bottom class for special study or for the purpose of correlating groups A, B, and C with an outside criterion, such a combination of manifest scale patterns, guided by the p values, seems quite justifiable. Of course, taking a cutting point like $p = .90$, as in the above example, is quite arbitrary. Any other cutting points, like $p = .95$ or $p = .67$, might be chosen, depending on the degree of "purity" desired in the extreme group. Even if we took $p = .50$ we would misclassify only about an eighth of our respondents.

We must recognize that in the case of the perfect latent dichotomy, with an indefinitely large number of manifest items, there are logically only two rank groups. Later, we shall describe latent structures *which logically postulate more than two ranks.* Nevertheless, it would be shortsighted, at least in the light of our present knowledge, not to use information of the type provided by Table 1, column 5, to order respondents into more than two empirical rank groups—particularly when our aim is to segregate those persons who are *highly likely* to possess the latent character or not to possess it.

The whole subject of the reliability of such groupings needs much more thorough theoretical and empirical study than it has been possible to make at the time these volumes go to press. In Chapter 10 Lazarsfeld has outlined the reliability problem and has suggested certain conditions which may justify expectations that, if the latent structure model is applicable, a relatively small number of items may be practically as serviceable as the much larger number of items conventionally employed in the absence of knowledge of the underlying structure.

For the ideal case, we can state analytically what is involved in the latent dichotomy approach by use of the following scheme for m items:

Latent class frequency	*Items* 1	2	3	4	.	*m*
n_{I}	p_{I1}	p_{I2}	p_{I3}	p_{I4}	. .	p_{Im}
n_{II}	p_{II1}	p_{II2}	p_{II3}	p_{II4}	.	p_{IIm}

Let the total number of respondents be n. Let the number of respondents favorable on item 1 be n_1, on item 2, n_2, etc. Let the number *jointly* favorable on items 1 and 2 be n_{12}, on items 1 and 3, n_{13}, etc. Let the number jointly favorable on items 1, 2, and 3 be n_{123}, on items 1, 2, 3, and 4 be n_{1234}, etc.

We seek to classify the population into two latent classes, one

class with frequency n_{I} and one class with frequency $n_{\mathrm{II}} = n - n_{\mathrm{I}}$. We need a set of $2m$ parameters $p_{\mathrm{I}1}, p_{\mathrm{I}2}, \cdots, p_{\mathrm{I}m}$ and $p_{\mathrm{II}1}, p_{\mathrm{II}2}, \cdots, p_{\mathrm{II}m}$. The value of $p_{\mathrm{I}1}$ is *the proportion, among all* n_{I} *respondents, who are favorable on item 1.* The value of $p_{\mathrm{II}1}$ is the proportion, among all n_{II} respondents, who are favorable on item 1, etc.

We must have:

$$
\begin{aligned}
n_{\mathrm{I}} + n_{\mathrm{II}} &= n \\
n_{\mathrm{I}}p_{\mathrm{I}1} + n_{\mathrm{II}}p_{\mathrm{II}1} &= n_1 \\
n_{\mathrm{I}}p_{\mathrm{I}2} + n_{\mathrm{II}}p_{\mathrm{II}2} &= n_2 \\
\cdots\cdots & \\
n_{\mathrm{I}}p_{\mathrm{I}m} + n_{\mathrm{II}}p_{\mathrm{II}m} &= n_m
\end{aligned}
$$

Now there must be no association between any two items among respondents belonging to a particular class. We have seen that, within class I, the proportion positive on item 1 is $p_{\mathrm{I}1}$ and the corresponding proportion positive on item 2 is $p_{\mathrm{I}2}$. Hence, if *within* class I, items 1 and 2 are to be independent, the proportion within class I who are jointly positive on items 1 and 2 must be $p_{\mathrm{I}1} \cdot p_{\mathrm{I}2}$. Similarly, the proportion within class I who are jointly positive on items 1 and 3 must be $p_{\mathrm{I}1} \cdot p_{\mathrm{I}3}$, etc. The same type of conditions must hold within class II—for example, the proportion within class II who are jointly positive on items 1 and 2 must be $p_{\mathrm{II}1} \cdot p_{\mathrm{II}2}$. Hence all relationships must be of the type: $n_{\mathrm{I}}p_{\mathrm{I}1}p_{\mathrm{I}2} + n_{\mathrm{II}}p_{\mathrm{II}1}p_{\mathrm{II}2} = n_{12}$.

Similarly, the proportions within class I of joint occurrence of positive references to items 1, 2, and 3 must be the product of the proportions within that class with positive responses to items 1, 2, and 3, that is, $p_{\mathrm{I}1} \cdot p_{\mathrm{I}2} \cdot p_{\mathrm{I}3}$; and analogously for all other joint occurrences involving any number of items.

The values of n_{I} and n_{II} and of the proportions $p_{\mathrm{I}1}$, $p_{\mathrm{I}2}$, etc., must be calculated from the data. When the latent dichotomy model does not fit the observed data perfectly as has been indicated, it can be made to fit an adjusted set of data. For the adjusted set of frequencies illustrated in Table 1, the values of the calculated parameters were:

Latent class frequency	*Items* 1	2	3	4
$n_{\mathrm{I}} = 424.3$	$p_{\mathrm{I}1} = .5125$	$p_{\mathrm{I}2} = .5704$	$p_{\mathrm{I}3} = .6276$	$p_{\mathrm{I}4} = .9240$
$n_{\mathrm{II}} = 575.7$	$p_{\mathrm{II}1} = .0635$	$p_{\mathrm{II}2} = .1008$	$p_{\mathrm{II}3} = .1871$	$p_{\mathrm{II}4} = .4324$

With these numerical values known one can compute straightforwardly the values shown in columns 3 and 4 of Table 1, by procedures shown in detail in Chapter 11.[10]

This exposition has used a single numerical illustration to help the reader grasp the basic idea of latent structure analysis, namely, that all the association between any two items is due to the latent attribute, and that there is no association between the specific components of the items. For the formal mathematics the reader is referred to Chapter 10. For the detailed explanation of how the adjusted frequencies are obtained when an approximate solution is necessary, the reader is referred to Chapter 11. In that chapter, each of the numerical steps in constructing Table 1 is shown in full, and the student may wish to repeat the operations step by step in order to appreciate more fully the assumptions and implications.

It should be said that certain of the problems of indeterminacy which haunt quantitative factor analysis appear in latent attribute analysis. Work with this new tool is still too young to have developed a completely standardized set of criteria for determining when the approximations made are "good enough" approximations. Such standards doubtless will be forthcoming, but they wait upon a large amount of further empirical and theoretical investigation. For example, two items which seemed by a priori inspection of content to be in the same general attitude area as the four items in our illustrative example were substituted for items 2 and 4. Although by conventional item analysis techniques these new items should belong in the scale, they did not satisfy the criteria for the latent dichotomy model.

As was said at the beginning of this section, the *latent dichotomy* model is merely a special case of the more general theory of *latent structure* analysis. Instead of two latent classes we may have λ latent classes in the following scheme:

Latent class	*Items*					
frequency	*1*	*2*	*3*	*4*	. .	*m*
n_I	p_{I1}	p_{I2}	p_{I3}	p_{I4}	. .	p_{Im}
n_{II}	p_{II1}	p_{II2}	p_{II3}	p_{II4}	. .	p_{IIm}
n_{III}	p_{III1}	p_{III2}	p_{III3}	p_{III4}	. .	p_{IIIm}
. . .	. . .	. . .	. . .	. . .	. .	. . .
n_λ	$p_{\lambda 1}$	$p_{\lambda 2}$	$p_{\lambda 3}$	$p_{\lambda 4}$	. . .	$p_{\lambda m}$

---

[10] For example, the number possessing the latent character among those with scale pattern + + + + in Table 1 is $n_I p_{I1} p_{I2} p_{I3} p_{I4} = 71.9$.

For the mathematical proofs one is referred to Chapter 10. The basic logic is exactly the same as in the latent dichotomy case, except that there are now λ latent classes instead of only two. The values of n_{I}, n_{II}, $\cdots$, n_λ and $p_{\mathrm{I}1}$, $p_{\mathrm{I}2}$, etc. are so determined that all the association between the responses to any two of the original items can be accounted for by the latent classes and by these alone.

It should be noted that unlike the latent dichotomy case, there is in the most general case no necessary intrinsic cue for *ranking* respondents. Only if the data possess certain special characteristics can a rank order be assigned to n_{I}, n_{II}, $\cdots$ and n_λ and hence a basis be provided for ultimately ranking respondents.

One of these special cases is of very great interest to us, however, since it corresponds closely to the scalogram case described in the first section of this chapter. When $\lambda = m + 1$, we may find that the quasi scale in scalogram analysis can be described as follows:

Latent class	*Items*					
frequency	*1*	*2*	*3*	*4*		*m*
n_{I}	$1 - a_1$	$1 - a_2$	$1 - a_3$	$1 - a_4$		$1 - a_m$
n_{II}	a_1	$1 - a_2$	$1 - a_3$	$1 - a_4$		$1 - a_m$
n_{III}	a_1	a_2	$1 - a_3$	$1 - a_4$	.	$1 - a_m$
.	.			. .	.	
n_m	a_1	a_2	a_3	a_4		$1 - a_m$
n_{m+1}	a_1	a_2	a_3	a_4	. .	a_m

Here $p_{\mathrm{I}1} = 1 - a_1$ where a_1 is a quite small fraction; $p_{\mathrm{I}2} = 1 - a_2$, where a_2 is also a small fraction; etc.[11]

As a_1, a_2, $\cdots$, a_m tend to zero, we approach the conditions for a perfect scale in the sense of scalogram analysis, and at the limit we have:

Latent class	*Items*				
frequency	*1*	*2*	*3*	*4*	*m*
n_{I}	1	1	1	1	1
n_{II}	0	1	1	1	1
n_{III}	0	0	1	1	1
.	.				
n_m	0	0	0	0	1
n_{m+1}	0	0	0	0	0

The theoretical implications of the fact that scalogram analysis may be shown to constitute a special case of latent structure theory need much further study.

[11] It is not necessary that all a's in a column, for example, be equal. But the special case in which they can be assumed to be very small and equal lends itself readily to computation because of the limited number of parameters involved.

In Chapter 11 a numerical example is worked out in detail applying latent structure analysis to data which form a quasi scale by the scalogram model. These are data on psychosomatic complaints in a sample of 1,000 soldiers. Four items are used in this illustration:

1. Have you ever been bothered by pressure or pains in the head?
2. Have you ever been bothered by shortness of breath when you were not exercising or working hard?
3. Do your hands ever tremble enough to bother you?
4. Do you often have trouble in getting to sleep or staying asleep?

The values of p in the structure, based on response to these four items, are as follows:

Latent class frequency	*Items* 1	2	3	4
n_{I}	.9640	.8703	.8473	.9326
n_{II}	.0360	.8703	.8473	.9326
n_{III}	.0360	.1297	.8473	.9326
n_{IV}	.0360	.1297	.1527	.9326
n_{V}	.0360	.1297	.1527	.0674

It is shown in Chapter 11 that when these values are used to distribute among the five latent classes the respondents who have any given response pattern (such as $+ + + +$) the results are very satisfactory. An intrinsic basis exists for ranking the response patterns and hence the respondents. Since approximations are involved, further study is needed as in the latent dichotomy case to set standards of acceptable approximations.

The scalogram model requires, of course, that the items tend to be cumulative. This restriction is also necessary when the latent structure analysis is applied to the scalogram picture as a special case. But there are other special cases of latent structure theory, which permit ordering of respondents without requiring the items to be cumulative. Consider the following picture:

Latent class frequency	*Items* 1	2	3	.	m
n_{I}	$p_{\text{I}1}{}^*$	$p_{\text{I}2}$	$p_{\text{I}3}$		$p_{\text{I}m}$
n_{II}	$p_{\text{II}1}$	$p_{\text{II}2}{}^*$	$p_{\text{II}3}$	.	$p_{\text{II}m}$
n_{III}	$p_{\text{III}1}$	$p_{\text{III}2}$	$p_{\text{III}3}{}^*$		$p_{\text{III}m}$
		.	.		.
n_m	p_{m1}	p_{m2}	p_{m3}		$p_{mm}{}^*$

Here the items and the latent class frequencies are so arranged that the largest values of the p's (denoted as p^*) are on the diagonal. *It is assumed that the values of* $p_{\text{II}1}, p_{\text{III}1}, \cdots, p_{m1}$ *diminish rapidly and monotonically from* $p_{\text{I}1}{}^*$. Similarly $p_{\text{I}3} < p_{\text{II}3} < p_{\text{III}3}{}^* > p_{\text{IV}3} > \cdots p_{m3}$, etc. This would appear to represent the type of phenomena present when items have been ordered on a Thurstone-type equal-appearing interval scale. Because of difficulties involved in computing the parameters, no numerical example is as yet available. One approach to the problem would be to cut down the number of parameters by imposing further symmetrical restrictions.

A particularly interesting special case of latent structure analysis is one of the cases where there are only three ordered latent classes. The algebra is such that arithmetic computation is comparatively simple, if we are willing to restrict the number of parameters involved in p to two per item. At the same time much flexibility in patterns of p values is possible. A given item could take on any of the following six patterns of p values:

	Latent class	*Value of p*		*Latent class*	*Value of p*
	I	$a_1 + b_1$		I	$a_4 + b_4$
(1)	II	a_1	(4)	II	$a_4 + b_4$
	III	a_1		III	a_4
	I	a_2		I	$a_5 + b_5$
(2)	II	$a_2 + b_2$	(5)	II	a_5
	III	a_2		III	$a_5 + b_5$
	I	a_3		I	a_6
(3)	II	a_3	(6)	II	$a_6 + b_6$
	III	$a_3 + b_3$		III	$a_6 + b_6$

As long as parameters are limited to two per item, computation difficulties are not substantially increased by increasing the number of latent classes. The values of a and b are free to vary from item to item and we are free to use any particular pattern or combination of patterns with any number of items. Such a pattern is illustrated in Chapter 11 with a numerical example from Research Branch data on job satisfaction.

A pattern such as one of the six above can be thought of as a discrete simplification of a continuous trace line $p(x)$ which is the function of a latent continuum x with an infinite number of classes. In introducing the concept of latent structure analysis, in Chapter 10, Lazarsfeld begins with the continuous case. He develops the theory for the continuous case and then considers a variety of special cases

in which there are a finite number of latent classes with varying patterns of p. Thus the six patterns given just above may be thought of as discrete approximations to curves of the second degree.

Among the most important contributions of Chapters 10 and 11 are preliminary tests enabling the investigator upon initial inspection of his data to decide which of various special cases would provide the more appropriate model for use on these particular data. These tests involve an examination of the first-order determinants formed from cross tabulation of each item with each other. The reader will find it particularly instructive to compare and contrast, in Chapter 11, the table of these determinants for the four items on attitude toward the Army which approximately fit a latent dichotomy pattern (page 419), with the set of determinants for the four items on psychosomatic complaints which are part of a quasi scale by the scalogram model (page 443). Each has its own pattern of internal consistency and the two patterns are strikingly different.

There are various additional theoretical and practical implications of latent structure theory, as outlined in Chapters 10 and 11. Latent dichotomy analysis has an especially interesting application to the problem of *shifts* in attitude by the same respondents at two time points. In Chapter 11 a numerical illustration from Research Branch data is given to show how the method serves in the analysis of shifts in attitudes toward officers.

As compared with the eight chapters devoted to scalogram theory the two chapters on latent structure analysis hardly do justice to the problems they raise. But the developments there reported are still too new to have associated with them the relatively large background of experience which is reported in the other chapters. In particular, computation techniques are usually much more difficult than with scalogram analysis and need simplification. Approximation criteria need further standardization—which is also true of some aspects of scalogram analysis.

Whether or not the latent structure theory lends itself readily to a large variety of practical applications, its generality and its suggestiveness as a model for the reduction of qualitative data make it worthy of much further exploration.

3. *Utilizing Indexes of Intensity of Attitude to Locate a "Zero Point" or "Region of Indifference"*

Thus far in this chapter we have previewed some of the contributions of this volume toward the problem of establishing among

respondents a rank order with respect to a particular attitude continuum.

In the last year or so of the Research Branch, progress also was made on the problem of establishing a "zero point" or "region of indifference" which would provide an objective basis, independent of particular question wording, for distinguishing relative proportions of pros and cons on an issue.

It developed that one of the most interesting theoretical properties of scalogram analysis was the fact that, in the ideal case of perfect scalability, the ranking of people with which we have thus far been concerned corresponds to but one of a set of *principal components*. In fact, as Louis Guttman proves in Chapter 9, if there are m response patterns or ranks there are m principal components. The conditions of perfect scalability, by scalogram analysis, lead to the result that each component score has a definite relation to the rank order of the people, and there is a definite law of formation.

The first component is a monotonic increasing function of the ranks (a straight line in the special case when the frequencies are the same for each response pattern). It is this first component with which we have been concerned in the problem of ordering people on a content scale.

The second component will always be a U-shaped or J-shaped curve with one bend in it. It need not be symmetrical, though in the special case where each response pattern has the same frequency it will be a symmetrical parabola. This second component has proved especially interesting, since it can be interpreted as a measure of intensity, leading theoretically, in the ideal case, to the determination of an objective zero point, at the point where the curve of intensity is a minimum.

The third principal component has two bends in it, the fourth three bends, and in general, the ith principal component has $i - 1$ bends. As yet little progress beyond the speculative stage has been made in investigating the psychological implications of the higher components, except for the third, which Guttman calls *closure*. (See Chapter 9, page 313.)

The first half of Chapter 9, in which Guttman introduces the theory of principal components, is written quite simply, for the nonmathematical reader. It is addressed to all psychologists and social scientists who are concerned with the problems of interpreting the socio-psychological implications of ordered structures. It will re-

pay the most careful study, as will, of course, the formal and rather difficult mathematical treatment in the latter part of the chapter.

The general idea of the first and second component will now be illustrated with a simple hypothetical example which has been developed more fully in Chapter 9. Consider the following scalogram, with all types of persons equally frequent and all types of item responses equally frequent:

Type of person (rank)	*Type of category*									
	Says yes to item					*Says no to item*				
	1	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*
5	X	X	X	X	X					
4		X	X	X	X	X				
3			X	X	X	X	X			
2				X	X	X	X	X		
1					X	X	X	X	X	
0						X	X	X	X	X
Frequency of response	1	2	3	4	5	5	4	3	2	1

Now we wish to assign a numerical *score* to each type of person and a numerical *weight* to each item response category. In the above example we want 6 person *scores* and 10 item *weights*.

The *scores* for persons should satisfy the following condition: All people who fall in one category of an item should have scores as similar as possible among themselves, and as different as possible from the scores of people in another category of the item. The total variance of persons' scores can be expressed as the sum of two parts: the variance of the scores *within* categories and the variance of scores *between* categories. Maximizing the similarities within categories and differences between categories implies maximizing the square of the *correlation ratio.*[12]

The *weights* for items should satisfy the following condition: All categories characterizing one person should have numerical weights as similar as possible, and as different as possible from the weights for categories which do not characterize that person. Maximizing similarities within people and differences between people calls for determining that set of numerical weights for all the item categor'es which will have the largest possible correlation ratio with respect to all the people.

Guttman has found that the maximum possible correlation ratio

[12] The correlation ratio is the square root of the ratio that the variance of scores between categories has to the total variance.

for scores is identical with the maximum possible correlation ratio for weights. He shows that the optimum score of a person is proportional to the arithmetic mean of the weights of the item response categories by which that person is characterized, and the optimum weight of a category is proportional to the arithmetic mean of the scores of the people who are in it.

The problem becomes one of finding a set of scores which "go in a circle"—that is, scores which yield weights that give back the same scores, differing from initial scores only by a constant of proportionality. Such a set of scores is called a principal component of the system. Let us now assign a set of *scores* to each rank type of person in the foregoing diagram as follows:[13]

5 to each response by a person in rank type 5
3 to each response by a person in rank type 4
1 to each response by a person in rank type 3
−1 to each response by a person in rank type 2
−3 to each response by a person in rank type 1
−5 to each response by a person in rank type 0

Then we have the following table:

Type of person (rank)	*Type of category* Says yes to item					Says no to item					*Sum of weights*	*Average score*
	1	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*		
5	5	5	5	5	5						25	5
4		3	3	3	3	3					15	3
3			1	1	1	1	1				5	1
2				−1	−1	−1	−1	−1			−5	−1
1					−3	−3	−3	−3	−3		−15	−3
0						−5	−5	−5	−5	−5	−25	−5
Sum of weights	5	8	9	8	5	−5	−8	−9	−8	−5		
Average weight	5	4	3	2	1	−1	−2	−3	−4	−5		

Now let us use the bottom row of numbers—the average weight for each item category—to compute a new score for each type of person. (For example, for a person with rank type 5 the weights $5 + 4 + 3 + 2 + 1 = 15$ or an average score of 3.) For the whole table we have:

[13] These scores are not arbitrary, but are the only set of scores (except for multiples thereof) which will satisfy the requirements of circularity for the first principal component. For details as to how these scores may be derived, the reader is referred to Chapter 9.

Type of person	*Type of category* *Says yes to item*					*Says no to item*					*Sum of weights*	*Average score*
	1	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*		
5	5	4	3	2	1						15	3.0
4		4	3	2	1	−1					9	1.8
3			3	2	1	−1	−2				3	.6
2				2	1	−1	−2	−3			−3	−.6
1					1	−1	−2	−3	−4		−9	−1.8
0						−1	−2	−3	−4	−5	−15	−3.0

Let us now compare the average score just obtained with that obtained in our initial trial:

Type of person	(x_1) *Average score on initial trial*	(y_1) *Average score on second trial (circular score)*
5	5	3.0
4	3	1.8
3	1	.6
2	−1	−.6
1	−3	−1.8
0	−5	−3.0

We see that the new sets of scores for types of persons is a linear function of the scores on our initial trial, namely $y_1 = .6x_1$. A continuation of this process will always lead to new scores which are exactly .6 times the previous scores. The set of x_1 scores we call the *first principal component.*

Now let us postulate and assign a *different* set of scores which also will be seen to go in a circle. The scores will be 5, −1, −4, −4, −1, 5. The ranks of these scores are no longer according to the scale rank order. Instead the scale rank score is "folded over." We have:

Type of person	*Says yes to item*					*Says no to item*					*Sum of weights*	*Average score*
	1	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*		
5	5	5	5	5	5						25	5
4		−1	−1	−1	−1	−1					−5	−1
3			−4	−4	−4	−4	−4				−20	−4
2				−4	−4	−4	−4	−4			−20	−4
1					−1	−1	−1	−1	−1		−5	−1
0						5	5	5	5	5	25	5
Sum of weights	5	4	0	−4	−5	−5	−4	0	4	5		
Average weight	5	2	0	−1	−1	−1	−1	0	2	5		

Entering the average weights from the bottom row, we have on our second trial:

Type of person	*Says yes to item*					*Says no to item*					*Sum of*	*Average*
	1	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*	*weights*	*score*
5	5	2	0	−1	−1						5	1.0
4		2	0	−1	−1	−1					−1	−.2
3			0	−1	−1	−1	−1				−4	−.8
2				−1	−1	−1	−1	0			−4	−.8
1					−1	−1	−1	0	2		−1	−.2
0						−1	−1	0	2	5	5	1.0

Comparing the average scores thus obtained with that which was obtained in our initial trial, we have:

Type of person	(x_2) *Average score on initial trial*	(y_2) *Average score on second trial (circular score)*
5	5	1.0
4	−1	−.2
3	−4	−.8
2	−4	−.8
1	−1	−.2
0	5	1.0

We see that the new set of scores for the type of persons is a linear function of the scores on our initial trial, namely, $y_2 = .2x_2$. A continuation of the process will always lead to new weights exactly .2 times the previous scores. The x_2 scores constitute the *second principal component.*

What we have been considering are only two of the principal components. Further analysis will reveal in the present example five principal components, that is, five sets of scores which have the property of circularity exhibited above.

Type of person	*Principal component scores*				
	I	*II*	*III*	*IV*	*V*
5	5	5	5	1	1
4	3	−1	−7	−3	−5
3	1	−4	−4	2	10
2	−1	−4	4	2	−10
1	−3	−1	7	−3	5
0	−5	5	−5	1	−1
Constant of proportionality	6	.2	.1	.06	04

It will be observed that the constants of proportionality (obtained through each respective circular operation) diminish with

each component. Since each constant of proportionality is, in fact, the square of the correlation ratio of a given component on the items, this means that each successive component accounts for a diminishing amount of the total variance in item categories. It also will be observed that when we add the five constants of proportionality we have

$$.6 + .2 + .1 + .06 + .04 = 1$$

that is, the sum of the squares of the correlation ratios is unity.

Principal components are, as Guttman observes, important concepts in mathematics and physics. Even though each item involves the same single factor as any other item, it differs in how it combines the components. "Each item represents," in Guttman's words, "a different variation on the same theme. Another way of expressing this is that each item represents a different combination of the tones and overtones of the same basic note." [14]

In Chapter 9, Guttman presents a detailed mathematical treatment of the problem and derives a number of further interesting properties of these functions. For example, he shows that though each principal component, in the ideal case, is a perfect curvilinear function of the rank order, the *linear* correlation between the principal components is zero.[15] He suggests that the comparatively simple law of formation as to their oscillations may prove to make the scalogram model of special theoretical importance among all the various models for ordering qualitative data. For scales which do not satisfy the scalogram criteria, the principal components, though always linearly uncorrelated as in the scalable case, have no simple pattern of dependencies and can, in fact, be completely independent. In general, there appears to be no thing, such as exists in the scalable case, as an inevitable U-shaped or J-shaped second component which might be identified as intensity, nor a third component which always has two bends, etc.

How important this theory of principal components will become, in enhancing the utility of scalogram theory as a tool for future investigation, it is much too early to say. It must be remembered that Chapter 9 deals with the ideal case, and perfect scalability by

[14] Other names which have been given in the literature to principal components include principal axes, latent vectors, characteristic functions, characteristic vectors, eigenfunctions, and eigenvectors.

[15] The reader can verify this in our simple numerical example above for any two components. For example, for II and IV we get $(5\cdot 1) + (-1\cdot -3) + (-4\cdot 2) + (-4\cdot 2) + (-1\cdot -3) + (5\cdot 1) = 0$.

the scalogram theory or any other theory is not likely to be found in practice.

Some months passed after Guttman's discovery of the principal components mathematically before a psychological interpretation of the second component was realized. The clue came from a paper by Daniel C. Katz in Hadley Cantril (editor), *Gauging Public Opinion.* Practical procedures for studying the intensity function were devised jointly with Edward A. Suchman, and it was Suchman who supervised the empirical research.

This empirical work is reviewed in Chapter 7, with several fully worked-out examples with Research Branch data. A few examples of intensity analysis were carried out in time to find a place in Research Branch studies which were not merely methodological. Illustrations appear in Volume I, Chapters 5 and 8 of the present series of books.

To find the intensity function it is necessary to develop, *independently* of the scale of content, a scale of intensity with which respondents felt about a subject. Two techniques were developed to arrive at such a scale of intensity. One was to follow each content item with a second item more or less like the following:

How strongly do you feel about this?
______ Not at all strongly
______ Not so strongly
______ Fairly strongly
______ Very strongly

Irrespective of the *direction,* pro or con, of the response on the preceding attitude item, a given intensity response, along with other intensity responses which were preceded by other attitude items, was used to form a scale of intensity. In no instance did the intensity items form a scale satisfying the criteria of scalogram analysis but they did ordinarily form a *quasi scale.*

An alternative method, which is called the *fold-over* technique, sought to derive the content and intensity scales from the same set of questions. For example, consider the item:

Do you think when you are discharged you will go back to civilian life with a favorable or unfavorable attitude toward the Army?
______ Very favorable
______ Fairly favorable
______ About 50–50
______ Fairly unfavorable
______ Very unfavorable

Such an item, as one of a set of items in an attitude scale, was now scored twice. First, it was scored for direction of content, in the usual scalogram manner. Second, it was scored for intensity—for example, the respondents who checked *either* top or bottom categories might be given a plus score and those who checked any of the three middle categories might be given a minus score. Intensity scores from such items were then combined to form a scale—actually, a quasi scale—of intensity. The fold-over technique was attractive, because it reduced the length of the questionnaire and relieved the respondent of a certain tedium. In a few studies it yielded about the same results as the longer method. But in other studies, it was less satisfactory. While the shorter procedure needs further study, the present disposition is to recommend the longer form wherever it is possible.

In Chapter 7 several numerical examples are provided of the empirical relationship between curves of intensity and scales of content for a variety of Army subject matter. In general, as predicted, a J-shaped or U-shaped curve was obtained with attitude data.

In the ideal case, the minimum point on the curve of intensity marks the zero point of the content scale. Empirically, however, there were a number of instances where the base of the U was broad and nearly flat, extending over several ranks. This phenomenon would seem to indicate a relatively broad "zone of indifference" and would make the selection of any single point as a minimum point somewhat hazardous, since the one point slightly lower than others might be so by chance. The sampling error of the minimum point has not been worked out and may not lend itself to a ready solution. Pending the development of better methods for designating a single point as a zero point in instances where the "zone of relative indifference" is broad, without a sharply defined minimum, one should treat the "zero point" with considerable caution. One procedure which has been suggested is to establish a "zone of indifference" arbitrarily by finding that rank at which each arm of the U-shaped curve has, say, median intensity. If a larger proportion of people with greater than median intensity are favorable than are unfavorable, this can be reported as a fact, along with the relative proportion favorable or unfavorable on either side of the "zero point." [16]

A crucial property of any satisfactory procedure for determining a

[16] This procedure could, of course, produce two contradictory conclusions. If so, and if the "zero point" is not sharply defined, the only safe conclusion may then be that the "zone of indifference" is broad, without specification of proportions favorable or unfavorable.

pro and con cutting point must be its independence of particular item wording.

In Chapter 7 it is reported, with numerical examples, that the methods there described seemed to have met this test quite satisfactorily. For example, a sample of six items from a scale of attitudes toward officers was used in which the proportion favorable on individual questions ranged from 8 to 43 per cent. This was called an *unfavorably biased* sample of items. Another sample of six items from a scale of attitudes toward officers was used in which the proportions favorable on individual questions ranged from 58 to 90 per cent. This was called a *favorably biased* sample of items. The respondents were the same on the two sets of items. It should be noted, parenthetically, that if one conventional public opinion polling agency had happened to use the first set of questions, whereas a second agency had happened to use the second set of questions, they would have come to opposite conclusions about the average proportion of men favorable to their officers. At least, they would have done so in the absence of some technique, *which was independent of particular item frequency,* for determining the proportions pro and con. For each of the two sets of items an intensity analysis was carried out. As is shown in detail in Chapter 7, approximately 80 per cent of the men were on the *unfavorable* side of the minimum intensity point, *irrespective of whether the favorably biased or unfavorably biased items were used.*

Further experimental work, to test the independence of the "zero point" or "region of indifference" from particular question wording, is needed on a wider variety of problems than the Research Branch had time to study in this way. If past experience continues to stand up in practice, there can be hardly any room for doubt about the utility of this device.

An interesting by-product of the empirical study of intensity functions has been the exploration of a concept of *generalized intensity.* This study came about through an effort to account for the rather wide dispersion of scores observed around intensity curves at all intervals.

Can it be, it was asked, that, irrespective of the particular attitude continuum under investigation, some people *habitually* are more vigorous in their responses than others? If so, their intensity of response to a set of items on a *particular* attitude continuum might need to be discounted. That is, the fact that they responded with apparent intensity to these attitude items might merely reflect general verbal habits and not necessarily strong feeling about the atti-

tude in question. The effect would be to increase the variability of responses around the intensity function in this or any other attitude area.

As described in Chapter 7, the problem was studied by constructing a quasi scale of generalized intensity, based on intensity of feeling expressed about a very wide variety of problems. The respondents were then broken into three groups, those with high, medium, and low generalized intensity, respectively. Selecting, for example, attitudes of men toward the Women's Army Corps (WAC), one computed intensity functions separately for these three groups of respondents. Each intensity function represented the relation between the content score of attitude toward the WAC and the specific intensity score belonging to this specific attitude area. The three functions had the same shape and *very nearly the same minimum point*, but the three functions formed a nest, with that for the men with high generalized intensity at the top and that for the men with low generalized intensity at the bottom. (See Chart I.) Necessarily, the variability around each of these three functions considered separately was less than around a single function formed by pooling all the cases irrespective of generalized intensity.

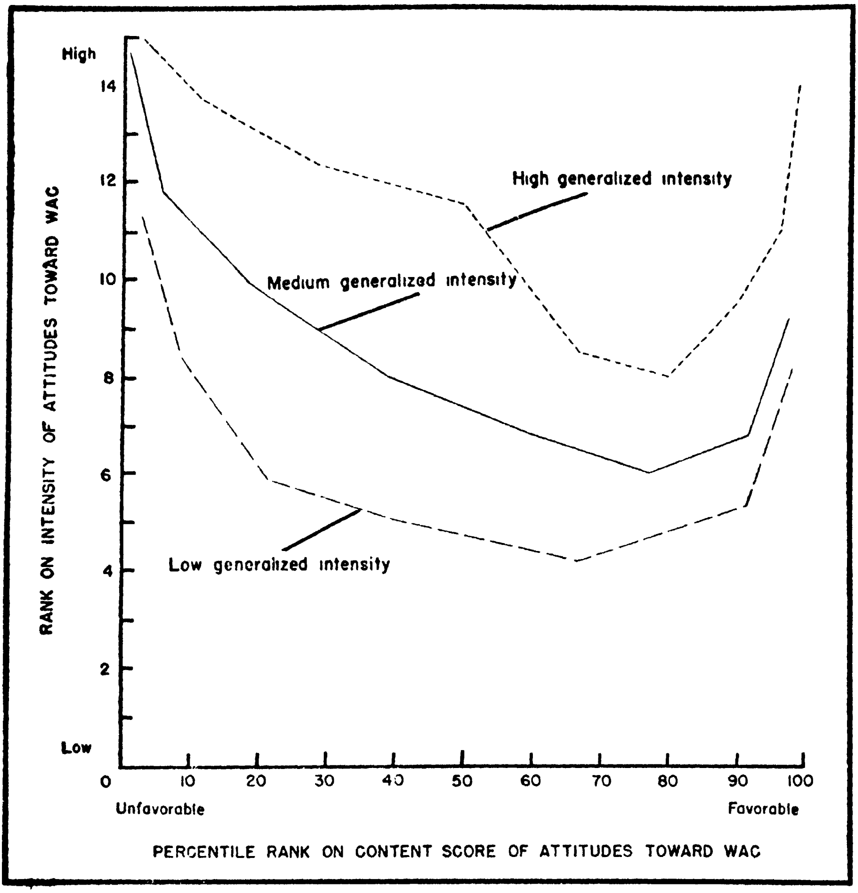
This type of exploration has important implications for attitude research in general and public opinion polling in particular, for it shows how responses to items involve mere verbal habits of expression which can lead to especially misleading results on individual items, as well as on scales. One person can say he feels very strongly opposed to something and yet be no more strongly opposed than another person who says that he feels somewhat opposed, the former being a person given to generally strong verbal habits and the latter to generally moderate verbal habits. The illustration given in the preceding paragraph showed why, if generalized intensity is not controlled, the variability around the specific intensity function will be rather large. It also gave encouraging intimations—which need to be tested on a very much wider variety of data—that even if the variability is considerable, the methods developed in the Research Branch will tend to yield a consistent "zero point" or "region of indifference" whatever the generalized level of intensity.

The implications of the theory of principal components and its empirical applications—especially, the second principal component, which in the ideal case is interpreted as the intensity function—need much further study. The ideas are not yet widely enough known to have received the kind of discerning criticism and new explora-

tion which is advancing understanding of those aspects of scalogram theory described in the first section of this chapter.

CHART I

SPECIFIC INTENSITY FUNCTIONS FOR RESPONDENTS AT THREE LEVELS OF GENERALIZED INTENSITY



Concluding Comments

The next ten chapters seek to present, as compactly as possible, due to present high printing costs, sufficient mathematical proof and numerical illustration to give the reader a good beginning in the understanding of the kind of thinking represented in the new approaches growing out of the work of the Research Branch.

Chapters 2 to 9 describe various aspects of scalogram analysis and Chapters 10 and 11 describe latent structure analysis. Most

of the chapters have been written primarily for the nonmathematician, the exceptions being parts of Chapters 9 and 10. All the illustrative numerical examples are from Research Branch data.

Each chapter is systematically organized and the attempt has been made to present the ideas as lucidly as possible consistent with rigor. Yet the reader is warned that these pages will require study and in places may be somewhat difficult, even where not primarily mathematical. Each chapter, of course, represents the point of view of its author and only a minimum editorial effort has been made to reconcile differences in terminology and in point of view, especially as between chapters written by Guttman and by Lazarsfeld, respectively.

In conclusion, it may be well to reiterate the point made at the beginning of this chapter, namely, that the purpose of this research is to throw new light on the problem of *constructing theoretical models of ordered structures or scales and testing the applicability of a particular model to a particular set of qualitative data.* Once we have found an appropriate model, we are in a position to operate with a relatively small number of items to describe a particular dimension, if the area is scalable by one or another rigorous criterion. Such short scales will be economical to use. When, in a given broad attitude region, such a set of quite specific short scales has been found, we should then be in a position to proceed with a multidimensional study of the area, testing theories as to the interrelationships of the scales used. This study may require us to try to reduce the number of these initial dimensions by some form of quantitative factor analysis. But the *first* step is to have a set of clean unidimensional scales and it is toward the attainment of this first step that Research Branch methodological energies were mainly directed.

As must be apparent from the preview presented in the present chapter, there is still relatively little which has sufficiently passed out of the realm of controversy to reach a definitive textbook stage. In the history of science, only a small fraction of the proposed scientific models, even including many which have certain initial attractions, find a permanent place. As more and more of the younger psychologists and social scientists, aware of the central importance of the problems here investigated, put their minds to investigation of these problems, we may expect models to take shape which will hold their place in science. Among those models may be some whose development has been furthered by the Research Branch and by the discussion, criticism, and creative inventiveness which it is hoped these chapters will stimulate.

CHAPTER 2

THE PROBLEM OF ATTITUDE AND OPINION MEASUREMENT[1]

The two problems. Two of the most important problems of all measurement concern (1) the determination of unidimensionality and (2) the determination of a fixed point of reference along such a single dimension. Measurements to be meaningful should be along only one dimension at a time. Once a single dimension has been isolated, it often becomes important to find some fixed point to which measures along this single continuum can be anchored.

The social sciences have long been concerned with both of these problems. Various techniques have been proposed for the determination of a single continuum, but, as will be discussed later, most existing tests for unidimensionality do not appear to be based upon any clear-cut rationale. With regard to the second problem of a fixed point of reference, little progress seems to have been made.

Our purpose is to offer a new theory and its concomitant techniques in answer to the two problems. The technique of *scalogram analysis*[2] provides a simple method for testing a series of qualitative items for unidimensionality, while the technique of the *intensity function* provides a simple method for finding a meaningful, objective cutting point along a single continuum.

The next chapters will discuss the problem of scalogram analysis, presenting in detail the basic theoretical foundations for scalogram analysis, a detailed outline of the scalogram board technique, illustrations of the utility of scale analysis in different problems, a discussion of the problems of reliability and validity for the present theory of scale analysis, and finally an evaluation of the present

[1] By Louis Guttman.

[2] Throughout the remainder of the chapters in this section, the terms scalogram analysis and scale analysis will be used interchangeably. The phrase "scale analysis" will refer only to the present technique of scalogram analysis.

theory of scale analysis as compared to existing methods for testing data to see if they lie along a single continuum.

After this discussion of scale analysis, the theoretical basis for the intensity function will be presented, followed by several illustrations of intensity curves and a demonstration of the independence of this curve and its cutting point from specific question wordings. Finally, the general equations of the components of scales will be derived, which in particular include the intensity component.

Scale analysis is formal. The problems of dimensionality and of an unbiased cutting point are of crucial importance in attitude and opinion research. Therefore, the present introductory chapter is devoted largely to a discussion of the definition and uses of attitude and opinion. The reader who wishes to proceed immediately to the theory and practice of scale and intensity analysis can omit the present chapter. The formal requirements for a scale hold regardless of what it is that is being studied, so that the principles of scale analysis hold in particular for any approach to attitude and opinion research.

Attitudes versus opinions? The fields of attitudes and opinions have often been thought of as presenting different problems in definition and measurement.[3] Differences in definition have been characterized by such terms as "controversial" (opinions) versus "noncontroversial" (attitudes), or "expressed" (opinions) versus "underlying" (attitudes). Differences in measurement have been

[3] There are several comprehensive summaries of attitude and opinion research to which the reader can refer for a more extensive treatment than is given here, especially in regard to specific surveys. Murphy, Murphy, and Newcomb present an excellent summary of specific surveys up to 1937; Schmeidler, Allport, and Veltfort summarize surveys during the war years; Sherif and Cantril present a detailed analysis and summary of studies on the formation and change of attitudes. Review of methodology may be found in Riker, Cantril, and especially in McNemar. This latter review and summary might well serve as an introduction to the problems studied in the present research.

G. Murphy, L. B. Murphy, and T. M. Newcomb, *Experimental Social Psychology* (Harper & Bros., New York, 1937).

G. W. Allport and H. R. Veltfort, "Social Psychology and the Civilian War Effort," *Journal of Social Psychology*, S.P.S.S.I. Bulletin, Vol. 18 (1943), pp. 165–233; G. R. Schmeidler and G. W. Allport, "Social Psychology and the Civilian War Effort: May 1943–May 1944," *ibid.*, Vol. 20 (1944), pp. 145–180.

Muzafer Sherif and Hadley Cantril, *The Psychology of Ego-Involvements* (John Wiley & Sons, Inc., New York, 1947). Also, Sherif and Cantril, "The Psychology of 'Attitudes': Part II," *Psychological Review*, Vol. 53, No. 1 (1946), pp. 1–25.

B. L. Riker, "A Comparison of Methods Used in Attitude Research," *Journal of Abnormal and Social Psychology*, Vol. 39, No. 1 (January 1944), pp. 24–42.

Hadley Cantril (editor), *Gauging Public Opinion* (Princeton University Press, Princeton, N.J., 1944).

Quinn McNemar, "Opinion-Attitude Methodology," *Psychological Bulletin*, Vol. 43, No. 4 (July 1946), pp. 289–374.

described in terms of "polling" (opinions) versus "scaling" (attitudes), or "dividing" (opinions) versus "ranking" (attitudes).

It seems generally agreed that "opinion" should refer only to verbal behavior, whereas an "attitude" may be either verbal or nonverbal. If we look at the research carried out on attitudes in the past, however, we find it is based largely on verbal behavior. There seems little point, then, in trying to distinguish between attitudes and opinions on this basis. We shall not attempt to define "opinion" as a distinct concept, but rather use the word interchangeably with "attitude" when dealing with verbal behavior. We shall use the word "attitude" to include "opinions" and also nonverbal behavior. When we use the phrase "attitude and opinion," this will be only to emphasize that both verbal and nonverbal characteristics are under consideration.

Polling versus scaling? The historical development of public opinion polls has followed somewhat different lines from what is ordinarily considered attitude research. Differences in particular techniques used by polling agencies from those developed for the general field of attitudes seem to have given rise to the impression in some quarters that opinion polls have different basic methodological problems from those of the general field.

Perhaps the outstanding difference is that opinion pollsters ask but a single question on a given issue, instead of a battery of questions as used generally in an attitude study. This is because pollsters are ordinarily interested in population percentages rather than in scoring each individual. Opinion polls cast their results into statements like "75 per cent of the population is favorable," or "35 per cent of the population prefers this." A more complete attitude study would want to say, "Person A is more favorable than Person B who in turn is more favorable than Person C. . . ."

If several questions are asked, then it is natural to inquire as to whether or not they are all tapping the same thing. Hence, the general concern of attitude students with the problem of scalability or dimensions of the responses. Is there a rank order for the people?

If but a single question is asked, then it is quite easy to succumb to forgetting about the problem of dimensions, but to worry primarily only about a cutting point. Is the question unbiased in its division of people into "favorable" and "unfavorable?"

A single question cannot provide any data to test for dimensionality, but this should not mislead one into believing that therefore the problem does not exist. If there is more than one dimension of

content in the question, what meaning can a cutting point have? On what is it that the population is "favorable" or "unfavorable?"

The problem of bias cannot really be tackled until it has been established whether or not a single dimension is present in the first place.[4] Hence, the problem of bias probably cannot be solved by the use of only a single question.. Public opinion polling is subject to the same basic problem of scalability as is the general field of attitude research.

In view of the foregoing considerations, it does not seem profitable methodologically to distinguish between opinion polling and other attitude research. The need for investigating dimensionality of responses remains in either case. Let us now turn to a closer view of the concept of attitude itself.

Definition in terms of observations. In this section is proposed an approach to a definition of attitude that is intended to promote research. *A complete definition is not attempted.* Only a necessary condition to be incorporated into a definition is suggested, namely, that an attitude be defined in terms of a delimited totality of behavior. Specifying behavior as a *necessary* condition can aid useful research without waiting for a complete definition.

A basic premise accepted here is that a scientific concept must be defined in terms of observations; it may be defined directly in terms of the observations, or by operations on the observations. A second premise is that a definition is scientifically useful only in so far as it leads to objective research. According to these premises, any sociological concept must be based on observations of human behavior and will be useful only to the extent that the requisite observations can be made and analyzed rigorously. In particular, these considerations pertain to the concept "attitude."

Many authors have proposed approaches to defining an attitude.[5] Two notions that seem most common to their discussions are that (a) an attitude is a predisposition to act in a certain way toward something (subject-object relationship), and that (b) it is an inference from previous behavior.

Let us examine these notions in the light of our basic premises. In order for them to provide a complete definition, the predisposition and the inference must be pinned down. The behavior to be pre-

---

[4] This undoubtedly accounts for the unrewarding results of the numerous subjective suggestions proposed by many writers on how to word questions so as to make them "unbiased."

[5] See summaries in the sources referred to in footnote 3.

dicted, the behavior from which inferences are to be drawn, and the process of inferring the predisposition must each be defined.

Prediction not sufficient. Merely to say that an attitude is that which helps predict certain behavior, of course, is not *sufficient* to define the term as conventionally used. It can be predicted that a person will behave in a certain manner, yet it would be agreed that the basis for the prediction should not be called attitudinal. For example, from a knowledge of the amount of education an enlisted man has had, we can predict to a certain extent how often he will go AWOL. But we would not say that a man's education should be an integral part of a *definition* of attitude toward Army regulations; it is a *correlate* and not a component. Relatively accurate predictions can be made of human behavior without direct reference to attitudes, and hence are insufficient to define an attitude.

Useful research on prediction requires that the predictor and the thing to be predicted must be defined separately and independently. The research problem is then the empirical one of finding out how well the predictor actually predicts one or more other variables.

Therefore, although how well an attitude can predict one or more other things will be of interest, the attitude cannot be defined merely as something that predicts. Two objections to such an approach are that (a) nonattitudinal variables also predict; and (b) to regard empirical correlates of an attitude as part of its definition makes nonsense of research. Prediction or correlation is not sufficient to define an attitude.

Similarly, prediction does not seem to be a *necessary* component of a formal definition of attitude. It may be that when attitudes are defined, they will yield useful predictions for some problems and less useful predictions for others. But if prediction is to be incorporated into the formal definition, then it must be stated how accurate it is. Must the prediction be perfect? If not perfect, and if predictability is measurable by correlation coefficients, is a correlation of .8 sufficient, or is .5, or is .01?

Perhaps even more crucially, if prediction is to be part of the definition, then the thing to be predicted—the criterion—must be defined. Shall the attitude of a person toward a candidate for election have as a criterion the voting behavior of that person on election day? Then those who are under-age, illiterate, or cannot vote for any other reason constitute a problem: do they or do they not have an attitude toward the candidate?

It seems implicit in conventional discussions of attitude as a pre-

disposition that behavior depends not only on the attitude but also on the situation. A person may hate his boss, but smile at him every day. Conceivably, then, attitudes may fail completely to predict some criteria.

If a definition of attitude could be made which involved prediction as a component, and which would facilitate research, this would be highly desirable. We believe that this has not yet been done, and that the problems raised in the preceding discussion militate against its being done. Attempting to define a predictor in terms of criteria to be predicted seems to lead to ambiguities and circularities. An alternative approach is to define attitude apart from any one criterion, and to regard prediction as an empirical correlate, rather than as a component. That is, after an attitude has been defined, the problem of prediction can then be tackled in a straightforward manner; the empirical correlations of the attitude with any outside variable desired can be ascertained. The attitude will be found to predict some things well in some situations, and some things poorly in some situations. In this way, many ambiguities are avoided.

We believe that in the long run better predictions will be made on the basis of research using such an approach. Preliminary evidence that predictions are facilitated by focusing on defining things in their own right is afforded by scalogram analysis (which is discussed in detail in the following chapters). In scalogram analysis, the universe of content is defined apart from any ideas of predicting outside variables. Then, if the content is scalable, it turns out that *any outside variable whatsoever* can have its maximum predictability determined in a very easy fashion.

The concept of an attitude universe. How can an attitude be defined without explicitly using the notion of prediction? An approach that seems to have possibilities of yielding a rigorous and useful definition is to consider an attitude to be a *delimited totality of behavior with respect to something.* For example, the attitude of a person toward Negroes could be said to be the totality of acts that a person has performed with respect to Negroes. A discussion of the problem of how to delimit such a totality is reserved for later.

Such an approach conforms to our premise for a definition by being based on observations. For the example of attitude toward Negroes, people can be observed with respect to how they behave toward Negroes in various kinds of situations. A person who votes against providing educational facilities for Negroes, sneers at Negroes when he sees them well dressed, speaks of them in deroga-

tory terms whenever he does talk about them, etc., could be classified as having one kind of attitude toward Negroes, and a person doing other things could be classified as having another kind of attitude toward Negroes.[6]

The utility of defining an attitude as a delimited totality of behavior depends on how complicated are the interrelationships between the various acts involved. It must be possible in practice to characterize the entire population whose attitude is being investigated on all the behavior involved in the definition. It should be possible to accomplish this either by (a) actual observation of each individual on each act; or (b) *by finding that the interrelations between the acts are such that from only a sample of acts the whole configuration can be reproduced.* If the number of acts is indefinitely large, a definition of attitude as a totality of acts can be useful only if the acts have a sufficiently simple pattern of interrelationships which will enable the whole to be reproduced fairly accurately from but a part. A scalogram pattern is one example of such a simple pattern of interrelationships.

One way of arriving at a workable definition in this sense is not to attempt to bite off too much in one chunk. For example, if a complete study of all behavior toward Negroes could be made, it might be found that this configuration was far too complex to be manageable as a whole, but that on the other hand, parts of the universe could be defined separately in a manageable form. The behavior of people toward Negroes when they meet them on the street might be proved to be a manageable configuration, or the behavior of people with respect to employing Negroes might be proved to be a manageable configuration; it would then be useful and profitable to speak of various attitudes: attitude toward Negroes when meeting them on the street, attitude toward employing Negroes, etc. The total attitude toward Negroes would then be the configuration of these subuniverses. A specific research project would rarely call for such a totality, but more often only for selected subuniverses.

Classifications and subclassifications. A totality or universe of behavior can usually be regarded as a subuniverse of a larger universe, and can itself often be divided into subuniverses.

A subarea of behavior is itself of interest if the larger area of which it is a part is of interest. Furthermore, the relationships of a subarea to outside variables can be studied and be useful regardless of

[6] Note the avoidance here of saying that one person is "more" favorable than another toward Negroes; it is not assumed that attitudes are necessarily scalable.

the role of the subarea in the total area. In an external prediction problem, if an adequate prediction can be made from only a subarea of an attitude, then that may be all that is needed for practical purposes; it need not be necessary to study the entire attitude.

By recognizing the different subuniverses that can be in an attitude, one can avoid the temptation of saying that a particular subuniverse is "really" the attitude. For example, it has been found that Negro soldiers will respond differently to questions put by Negro interviewers than to the same questions put by white interviewers. Instead of asking which situation elicits the "real" attitude of the respondents, it seems far more profitable to recognize that here are two distinct, though related, subuniverses. If it is desired to predict something external from these responses, it may be found that the responses involving the Negro interviewer will help predict how a respondent will act in the company of Negroes with respect to the issue at hand; whereas, the responses elicited by white interviewers may better predict how the Negro will behave in a white environment. And most certainly, knowledge of the behavior in *both* the white and Negro interview situations will in general provide better external predictions than will knowledge of only one of the subuniverses.

The role of questionnaire research. An interesting feature of current research is the use of questions in stimulating responses. The behavior performed by a respondent on a written questionnaire is either to place a check mark opposite a category of the printed answers to a printed question, or to write out a response by himself. In an interview situation, this writing is done by the interviewer himself. If the content of the questions is, for example, expression of opinion about the respondents' officers, then the responses are classifiable as a subuniverse of the men's attitude toward their officers.

Such a subuniverse of behavior of responding to a questionnaire is, of course, not the same as other behavior with respect to officers; and the relationships it has with other subuniverses of the attitude is a matter for empirical investigation, as is any question of interrelationship between subareas.

Questionnaire responses at present are by far the most manageable kind of subuniverse of attitudinal behavior. If the attitudes of a population are under investigation and if it is not possible to observe the entire population on the totality of acts, then there are two kinds of sampling problems involved. One is the sampling of

acts, and the other is the sampling of people. Although the universe of verbal behavior observable by the questionnaire technique is indefinitely large, an adequate sample can probably be incorporated into an administrable questionnaire, and a proper sample of people can also be obtained.[7] Many items can be included in a questionnaire and responded to in a very short period of time, whereas other kinds of subuniverses of behavior may take place over long periods of time and be difficult to observe systematically for a satisfactorily large group of people.

Like any other subuniverse, questionnaire responses can be used to predict outside variables if desired, the amount of predictability being a matter to be ascertained empirically. If sufficient predictability can be obtained from such a subuniverse alone, then there may be no need to investigate the other subuniverses of the attitude for that particular problem.

The insufficiency of "known-group validation." Is it true that the attitude-as-behavior approach can lead to improvement in research? Let us examine how it bears on two conventional techniques. One can be called that of *known-group validation*, and the other that of the *single criterion.*

The procedure of "known-group validation" consists of beginning with an informal judgment of one or more groups of people with respect to their attitude. The behavior defining the attitude is not explicitly stated beforehand, but it is agreed to accept comparisons of the groups with respect to this undefined attitude. An illustration might be that of "esprit de corps." The "known-group" approach for this is to say, for example, "We don't know exactly what the term signifies, but we can agree that whatever 'esprit de corps' is, the paratroops in the Army have more of it than the Infantry. Therefore, let us try to find observations that will best discriminate between paratroopers and infantrymen."

A patent weakness of such an approach is that it cannot provide a definition. Many things could be found that would discriminate between paratroops and Infantry, but would not be called "esprit de corps." As an extreme example, the shoes men wore (at least until late in the war) would discriminate very sharply between paratroops and Infantry; but a consensus of research workers would undoubtedly reject shoes as part of the definition of "esprit de corps."

---

[7] The processes of sampling people and of sampling items are not at all identical; random sampling, stratified or not, is used for the first, but is not applicable to item construction.

This approach cannot distinguish between a definition and its correlates.

A less obvious weakness in this indirect technique is its implicit assumption that an attitude is necessarily unidimensional, that individuals or groups can be meaningfully ranked on it from high to low. If an attitude is composed of several multidimensional elements, there is no need for each of the separate elements to discriminate between groups even though the composite does discriminate. It is a commonplace of multiple correlation that zero-order relationships may have little bearing on higher order relationships.

Such an indirect validation approach is not designed for defining concepts; it seems intended only to help in the absence of a definition, and what help it can give is necessarily very limited. Criterion groups can be useful as a stimulation to intuition in cases where it is not clear what the problem is. By asking, "What behavior led us to think of these groups as criterion groups?" research workers can be led to define one or more classes of acts that can be agreed on as being the behavior the researcher is interested in studying. From the point of view of attitude-as-behavior, the important thing is finally to be clear as to the acts that are to be considered intrinsic to the attitude. In so far as the criterion groups prove to clarify the universe of interest, they are an effective informal aid; a danger is to focus so much on the groups as to lose sight of the fact that possible discrimination between them is only a correlate of the attitude or attitudes to be studied.

Dangers of the regression approach. A related indirect approach widely used is that of attempting to predict a single criterion. As an example, consider the notion of job satisfaction. The approach of the single criterion requires that the attitude of job satisfaction—when defined—should correlate as highly as possible with actually leaving or staying on the job when it is a matter of free choice (or some similar criterion). Things that could help predict whether or not a person will leave a job, given his free choice, would be assembled into a composite to predict the criterion; and the regression of the criterion on the composite would be called an *index* of job satisfaction. The attitude itself is not defined, but the index is supposed, in some sense, to *represent* the attitude.

For the purpose of predicting whether or not a man will leave his job, the procedure of assembling a composite and coming out with a prediction variable is, of course, the proper procedure. It should not, however, be inferred that thereby a concept is necessarily de-

fined. How this would be an improper inference can be seen in several ways.

One way is to suppose that a composite can be assembled that would yield a prediction variable that allows perfect prediction. There is a curious aspect to such a composite, namely, that it is *never unique*. If more variables are added to the composite, the enlarged composite will necessarily also yield perfect prediction, *no matter what the additional variables are.* This is an immediate consequence of the definition of multiple regression; adding variables can not reduce predictability. Furthermore, in general it should not be expected that there is a unique smallest subset of variables in the composite that will yield perfect prediction. It seems plausible that two or more different subsets of predictors will yield the same perfect prediction, if perfect prediction is possible at all.

Similarly, in the case where a composite yields imperfect prediction—which is the usual case in practice—there are in general several different subsets (or several other composites) which will yield equally good predictions.[8] Furthermore, for this case of imperfect prediction, there can be negligible correlations between the indexes.[9]

If it is desired to predict some external variables from the attitude, and if all that is available is an index which correlates but imperfectly with the attitude, it is very unsatisfactory merely to correlate the external variable with this index. Which of the possible indexes should be used? Even though the indexes should correlate equally with the attitude, their correlations with an external variable can vary considerably, just as the correlations among themselves can vary considerably. *It is possible for an index to predict the outside variable far better than does the attitude, as well as far worse than does the attitude.*

And even further, the correlation of an external variable with a particular index is in general less than the multiple correlation of the external variable on the observations from which the index was derived. The multiple regression of the attitude on the prediction composite is in general different from the multiple regression of an external variable on the same composite, so that the regression for

---

[8] For an analytical discussion of the lack of intrinsic meaning of multiple regressions, see sections 19, 20 and 21 in Louis Guttman, "An Outline of the Statistical Theory of Prediction," in P. Horst, et al., *The Prediction of Personal Adjustment* (Social Science Research Council, New York, 1941), pp. 287–292.

[9] As a numerical example from linear correlation, consider the case where each of two composites correlate .7 with the criterion. It can be proved that the only restriction on the correlation between the two indexes is that it be between $-.02$ and $+1.00$.

the one variable is not the most appropriate to be used for the other.

In view of all this, it seems that defining an attitude by resorting to an index can lead only to ambiguous research.

The characterization of attitude as a delimited totality of behavior is intended to de-emphasize the indirect approaches which characterize so much of current studies. The ambiguities of indirection are almost of necessity irresolvable. A direct approach such as outlined here seems better designed to clarify and to structuralize research problems into a usable form, and the problems that attend studying attitude-as-behavior have the possibility of being solved.

Internal validity and external validities. Perhaps the major reason why so much research has been predicated upon indirect and incomplete validation procedures has been lack of recognition of the fact that in general there are two quite distinct kinds of validity. One kind we shall call "internal" and the other kind "external."

"Internal validity" is the problem of definition. In defining a universe of behavior, the test of "internal validity" for each item in the universe is its content. Does the content of the item belong in the universe? Only a judgment of the content can answer this, according to our approach. For example, does the following problem belong to a universe of knowledge of arithmetic: "How much is two plus two?" Or does the following item belong in a universe of opinion about the presidential qualities of Mr. Truman: "Do you think Truman is doing a good or a bad job as President?" If one wishes to study knowledge of arithmetic, then one asks arithmetic questions. If one wishes to study opinions about the President, then one asks questions about the President. What questions go into each of such universes is determined by the content involved. Each item with the proper content of the universe has "internal validity" for that universe.

"External validity" is the problem of prediction. A universe has but one "internal validity," but *it has many possible "external validities,"* since it can be used for many different prediction purposes. Knowledge of arithmetic may have some validity for predicting success as an accountant, a different validity for predicting success as an engineer, etc., etc. Opinion of President Truman may have some validity in predicting who will vote for him on election day, a different validity for predicting who will support his program on price control, etc., etc. A particular "external validity" for a universe cannot be determined except by experiment. The thing to

be predicted must first be defined, and then observations must be made to determine the correlation between the predictor and this particular criterion. Generally, the correlation will not be perfect. In the social sciences, we rarely find that one variable can be predicted well from another. Ordinarily, a battery of predictors is needed to attain a substantial correlation with a criterion. To predict success as an accountant, it should be helpful to consider not only knowledge of arithmetic, but other universes as well. To predict well who will vote for a presidential candidate, other things besides opinion of that candidate should be taken into account. Hence, the many "external validities" of an attitude may be far from perfect by themselves.

The "external validities" of an attitude cannot be studied until the attitude itself is defined. "Internal validity" must be agreed on first. After the attitude is defined, then any particular "external validities" can be explored empirically. Some may be high and some may be low, but, regardless, it should be clear that "external validities" are empirical correlates of, and not integral criteria for, "internal validity."

Accuracy as an "external validity." The distinction between "internal" and "external validity" can be illustrated by one kind of prediction problem, that of *accuracy*. In a poll or survey, some questions may be used which attempt to reproduce some external or objective fact, and some questions may be used to elicit opinion.

If a man is asked to state his age, one "external validity" of his response would be the accuracy with which it corresponds with data recorded on his birth certificate. This particular external criterion on the birth certificate is defined apart from the man's response, and can be observed without ever asking the man any questions. The man's response is a separate variable which belongs to a universe of all questions that could be asked of the man about his age; it has "internal validity" for this universe of questions because of its content.

What a man says about his age is one thing, and what the birth certificate records is another; the relationship between the two is something that is to be determined empirically. But, regardless of this relationship, regardless of this particular "external validity" of the man's response, the response can have high "external validity" for other prediction problems. Even if the response does not accurately represent birth certificate data, it is a statement of the man about himself, and such a statement can quite often predict other

things very well. A man's statement about his age might not be accurate chronologically, but it might prove to be a good predictor, say, of his personal adjustment or other areas defined externally to it. Therefore, to say that a man's statement about his age is "not valid" is very incomplete; what is ordinarily meant by this is that it has little "external validity" for one particular problem.

In opinion questions, there is less danger of losing focus by seizing upon a particular external criterion, because there ordinarily is no supposedly factual counterpart to the opinion response which can be obtained independently of the respondent. To inquire into the "accuracy" of opinions does not in general make much sense, then. Perhaps what is referred to by such an inquiry are problems concerning the sampling of questions and of bias. How well does a particular opinion question represent the universe from which it was drawn? Does it divide the population properly into pros and cons? Such problems are the concern of scale and intensity analysis, and are *internal* to the universe. They are not problems of "external validity."

The following chapters are devoted to the study of the internal structure of a universe of attitude and opinion items (or of any other universe of qualitative data), and show how knowledge of this structure is essential for an analysis of such qualitative data.

CHAPTER 3

THE BASIS FOR SCALOGRAM ANALYSIS[1]

One of the fundamental problems facing research workers in the field of attitude and public opinion measurement is to determine if the questions asked on a given issue have a single meaning for the respondents. Obviously, if a question means different things to different respondents, then there is no way that the respondents can be ranked in order of favorableness. Questions may appear to express but a single thought and yet not provide the same kind of stimulus to different people. The responses even to the simplest question can differ in kind as well as in degree.

That two people can give the same response to the same question and yet have different attitudes has long been of primary concern to public opinion pollsters. This is particularly the case when more than one question is asked about the same topic, and the replies to the different questions appear to be inconsistent. Consistency of response is a problem that plagued attitude research long before public opinion polling captured the fancy of the public. How can one tell if there is enough consistency in the responses of a population to a series of questions to indicate that only a single factor is being measured? Are all respondents interpreting the questions to mean the same things? Are differences in responses due only to differences in degree of feeling and not to differences in kind? Is it meaningful to score the people from high to low with respect to a given set of items?

This problem of consistency underlies a great deal of research in the social and psychological sciences dealing with large classes of qualitative observations. For example, research in marriage is concerned with a class of qualitative behavior called marital adjustment which includes an indefinitely large number of interactions between husband and wife. Public opinion research is concerned with large

[1] By Louis Guttman. A bibliography of published articles on scalogram analysis is given at the end of this chapter.

classes of behavior like expressions of opinion by Americans about a military treaty with the British. Educational psychology deals with large classes of behavior like achievement tests. Other problems in social and psychological research where the consistency underlying a series of items is of fundamental importance include the study of aptitudes involved in fitting people into jobs, the measurement of human intelligence and abilities, the study of neurotic behavior and other aspects of personality, the appraisal of social status—in short, any problem involving the assigning of numerical values to qualitative observations in an attempt to evolve a single rank ordering. It is often desired in such areas to be able to summarize data by saying, for example, that one marital couple is better adjusted than another marital couple, or that one person has a better opinion of the British than has another person, or that one student has a greater knowledge of arithmetic than has another student.

While the data to be presented in the present chapter will deal almost entirely with social attitudes or opinions, the approach is appropriate for all of the types of problems mentioned. This rather new approach seems to afford an adequate basis for the quantification of many types of qualitative data.

The approach to scale analysis to be presented here has been used successfully during the war in investigating morale and other problems in the United States Army. While some interesting mathematics lie in the background, no knowledge of this mathematics is required in actually analyzing data. Simple routines have been established which require no knowledge of statistics, which take less time than the different manipulations now used by various investigators (such as critical ratios, biserial correlations, factor analyses, etc.), and which give a complete picture of the data not afforded by any of these other techniques. The word "picture" might be interpreted here literally, for the results of the analysis are presented and easily assimilated *in the form of a "scalogram," which gives the configuration of the qualitative data.*

Theory of Scale Analysis

A main condition for an attitude scale has often been pointed out by psychologists. For example, Murphy, Murphy, and Newcomb say that "no scale can really be called a scale unless one can tell from a given attitude that an individual will maintain every atti-

tude falling to the right or to the left of that point." [2] A similar consideration is the starting point for our theory of scale analysis. Instead of focusing on the ranking of items, however, we focus on the ranking of individuals. The ranking of items apparently is restricted to dichotomous items, where a person either endorses or does not endorse a statement. In such a case, it is possible to consider a ranking of endorsements, so that if a person endorses a more extreme statement, he should endorse all less extreme statements if the statements are to be considered a scale. But if the items have more than two categories, such a consideration breaks down; "agree" to one item might be equivalent to, or even less "favorable" than, "undecided" to another item, so that there remains a problem of how to rank items and response categories.

The ranking of people provides a more general approach to the problem of scaling, since it turns out to be equivalent to the ranking of items when all items are dichotomous, and it also includes the case where items have more than two answer categories. We shall call a set of items of common content a scale if a person with a higher rank than another person is just as high or higher on every item than the other person. This involves no problem of ranking the categories of one item against those of the other items, but only needs a ranking of the categories *within* each item.

An equivalent definition of a scale for our approach is that, within each item, if one response category is higher than another, then *all* people in the higher category must have higher scale ranks than those in the lower category.

A third equivalent definition of a scale is the one upon which our practical scalogram analysis procedures are directly based. It requires that each person's responses should be reproducible from his rank alone. A more technical statement of the condition is that each item shall be a simple function of the persons' ranks. The meaning of this is expanded below, where it is easily seen that the three definitions of a scale just given here are all equivalent. This third definition, while perhaps the least intuitively obvious, has proved to be the most convenient formulation for practical procedures. Of course, when it is fulfilled, the first two definitions are also fulfilled; and when the items are dichotomous, then the psychologists' condition first mentioned is also fulfilled.

---

[2] G. Murphy, L. B. Murphy, and T. M. Newcomb, *Experimental Social Psychology* (Harper & Bros., New York, 1937), p. 897. We shall use the word "item" where they use "attitude."

Murphy, Murphy, and Newcomb further note that, "As a matter of fact there is every reason to believe that none of the rather complex social attitudes which we are primarily discussing will ever conform to such rigorous measurement." [3] Perhaps such a belief may account for the fact that the mass of current attitude research pays little or no attention to this fundamental rationale. The common tendency has been to plunge into analysis of data without having a clear idea as to when a single dimension exists and when it does not. For example, *bi*variate techniques—like critical ratios and biserial correlations—are commonly used to find items that "discriminate" and to determine "weights," without testing whether or not the *multi*variate distribution of the items is actually indicative of a single dimension.

One of the main purposes of this chapter will be to propose a rigorous definition of a scale. This definition applies not only to the study of general attitudes, but also to the study of public opinion on specific issues. Furthermore, specific examples will be given to show that *consistent scales satisfying the rigorous requirements proposed above have been obtained in actual practice.* The proposed method provides a simple analysis of a series of questions which will enable one *to determine quickly whether or not the basic condition for a scale is satisfied by the data.*

The notions of variable, function, and simple function. First, a word about what is meant by a variable, whether qualitative or quantitative. We use the term in its conventional logical or mathematical sense, as denoting a set of values. These values may be numerical (quantitative) or nonnumerical (qualitative). We shall use the term "attribute" interchangeably with "qualitative variable." The values of an attribute (or of a quantitative variable, too, for that matter) may be called its *subcategories*, or simply *categories*.

A variable y is said to be a single-valued function of a variable x if to each value of x there corresponds a single value of y.

In particular, suppose y is an attribute, say like the attribute about expression of liking for the British, and takes on three values. We may denote by y_1 the statement, "I like the British"; by y_2, the statement, "I don't like the British"; and by y_3, "I don't know whether or not I like the British." If x is a quantitative variable which takes on at least three values, and if we can divide the x values into three intervals which will have a one-to-one correspond-

---

[3] *Ibid.*

ence with the values of y, then we shall say the attribute y is a *simple* function of x. For example, suppose x takes on the ten values 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Then the correspondence table might be as follows:

x	0	1	2	3	4	5	6	7	8	9
y	y_1	y_1	y_1	y_3	y_3	y_2	y_2	y_2	y_2	y_2

The three intervals for x are: 0–2, 3–4, and 5–9, to which correspond the values y_1, y_3, and y_2 respectively. Every person who has an x-value between 0 and 2 has y_1 as his y-value; every person who has an x-value of 3 or 4 has y_3 as his y-value; and every person with an x-value between 5 and 9 has y_2 as his y-value.

We might show this graphically by plotting the x-values on a straight line, and cutting it into intervals:

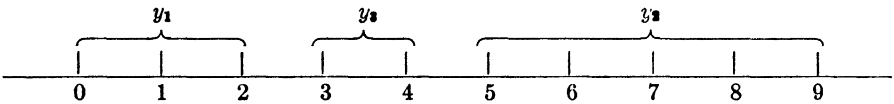


Figure 1

For statistical variables, another representation is in terms of a bar chart of frequencies, and this is what is used for convenience below.

The definition of scale. For a given population of objects, the multivariate frequency distribution of a universe of attributes will be called a *scale* if it is possible to derive from the distribution a quantitative variable with which to characterize the objects such that each attribute is a simple function of that quantitative variable. Such a quantitative variable is called a scale variable.

Perfect scales are not to be expected in practice. The deviation from perfection is measured by a *coefficient of reproducibility*, which is simply the empirical relative frequency with which the values of the attributes do correspond to the proper intervals of a quantitative variable. In practice, 90 per cent perfect scales or better have been used as efficient approximations to perfect scales.

A value of a scale variable will be called a *scale score*, or simply a *score*. The ordering of objects according to the numerical order of their scale scores will be called their scale order.

Obviously, any quantitative variable that is an increasing (or decreasing) function of a scale variable is also a scale variable. For

example, in the illustration above, consider x to be a scale variable. Any constant could be subtracted from or added to each of the x scores, and y would remain a simple function of the transformed x. Thus, the scores 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 could be replaced by respective scores -5, -4, -3, -2, -1, 0, 1, 2, 3, 4. Or the x scores could be multiplied by any constant, or their square roots or logarithms could be taken—any transformation, continuous or discontinuous, could be used, as long as the rank order correlation between the original x and the transformed variable remained perfect. All such transformations will yield scale variables, each of which is equally good at reproducing the attributes.

Therefore, the problem of metric is of no particular importance here for scaling. For certain problems like predicting outside variables from the universe of attributes, it may be convenient to adopt a particular metric like a least squares metric, which has convenient properties for helping analyze multiple correlations. However, it must be stressed that such a choice of metric is a matter of convenience; any metric will predict an outside variable as accurately as will any other.

In practice, the rank order has been used as a scale variable. (It is in fact a least squares metric for a rectangular distribution of scale scores.)

While rank order is sufficient for the *mechanical* aspects of testing for scalability and of external prediction, it is not adequate for certain problems of *psychological* description and generalization. The equations of scale analysis in Chapter 9 provide a psychologically meaningful metric, beyond rank order, whose zero point is located by means of intensity analysis as in Chapter 7.

An Example of a Scale of Dichotomies

As may be expected, the universe of attributes must form a rather specialized configuration if it is to be scalable. Before describing a more general case, let us give a little example. Suppose that a statistics test is composed of the following problems:

Consider a population of voters in which 60 per cent are Democrats and 40 per cent are Republicans.

1. What is the probability that one person chosen at random will be a Democrat?
2. What is the probability that two people chosen at random will both be Democrats?

3. What is the probability that out of ten people chosen at random, *at least* three will be Democrats?

If this test were given to the population of members of the American Sociological Society, we would perhaps find it to form a scale for that population. The responses to each of these questions might be reported in dichotomous form as "right" or "wrong." There are $2 \times 2 \times 2 = 8$ possible types for three dichotomies. Actually, for this population of sociologists we would probably find only four of the eight types occurring. There would be (a) the type which would get all three questions right, (b) the type which would get the first and second questions right, (c) the type which would get only the first question right, and (d) the type which would get none of the questions right. Let us assume that this is what would actually happen. That is, we shall assume the other four types, such as the type getting the first and third questions right but the second question wrong, would not occur. In such a case, it is possible to assign to the population a set of numerical values like 3, 2, 1, 0. Each member of the population will have one of these values assigned to him. This numerical value will be called the person's score. From a person's score we would then know precisely to which problems he knows the answers and to which he does not know the answers. Thus a score of 2 does not mean simply that the person got two questions right, but that he got two particular questions right, namely, the first and second. A person's behavior on the problems is reproducible from his score. More specifically, each question is a *simple function* of the score, as will be shown below.

The meaning of "more" and "less." Notice that there is a very definite meaning to saying that one person knows "more" statistics than another with respect to this sample. For example, a score of 3 means more than a score of 2 because the person with a score of 3 knows *everything a person with a score of 2 does, and more.*

There is also a definite meaning to saying that getting a question right indicates "more" knowledge than getting the same question wrong, the importance of which may not be too obvious. People who get a question right all have higher scale scores than do people who get the question wrong. As a matter of fact, we need no knowledge of which is a "right" answer and which is a "wrong" answer beforehand to establish a proper order among the individuals. For convenience, suppose the questions were given in a "true-false" form, with suggested answers (1) .50, (2) .36, (3) .42 for the respec-

tive questions.[4] Each person records either a T or an F after each question, according as he believes the suggested answers to be True or False. If the responses of the population form a scale, then we do not have to know which are the correct answers in order to rank the respondents (only we will not know whether we are ranking them from high to low or from low to high). By the scale analysis, which essentially is based on sorting out the joint occurrences of the three items simultaneously, we would find only four types of persons occurring. One type would be $F_1T_2F_3$, where the subscripts indicate the questions; that is, this type says F to question 1, T to question 2, and F to question 3. The other three types would be $F_1T_2T_3$, $F_1F_2T_3$, and $T_1F_2T_3$. These types can be shown in a chart (a "scalogram") where there is one row for each type of person and one column for each answer category of each question.

The scale analysis would establish an order among the rows and among the columns which would finally look like this:

			Question			
Type score	F_3	T_2	F_1	T_3	F_2	T_1
3	x	x	x			
2		x	x	x		
1			x	x	x	
0				x	x	x

Figure 2

Or, alternatively, both rows and columns might be completely reversed in order. Each response to a question is indicated by an x. Each row has three marks because each question is answered (either correctly or incorrectly). The *parallelogram* pattern in the chart[5] is necessary and sufficient for a set of *dichotomous* attributes to be expressible as simple functions of a single quantitative variable.

From this chart we can deduce that F_1, T_2, and F_3 are all correct answers, or are all incorrect answers. That is, if we were now told that F_1 is a correct answer, we would immediately know that T_2 and F_3 are also correct answers. This means that we can order the men according to their knowledge even if we do not know which are

[4] We shall assume that no one gets an answer right by guessing. Scale analysis can actually help one pick out responses that were correct merely by guessing from an analysis of the pattern of errors. But for this, much more than three items are necessary.

[5] Such a chart, where one column is used for each *category* of each attribute, we call a *scalogram*. The scalogram boards used in practical procedures are simply devices for shifting rows and columns to find a scale pattern, if it exists, as will be explained in detail in the next chapter.

the correct answers and which are the incorrect answers, only we do not know whether we are ordering them from highest to lowest or from lowest to highest. Except for direction, the ordering is a purely formal consequence of the configuration of the behavior of the population with respect to the items. The importance of this fact will become more apparent in more complicated cases where the attributes are not dichotomous but have more than two categories. As will be shown later, the scale analysis automatically decides, for example, where an "undecided" response to a public opinion poll question belongs, whether it is above "yes," below "no," in between, equivalent to "yes," or equivalent to "no." A priori judgments of content order are not essential to scale analysis.

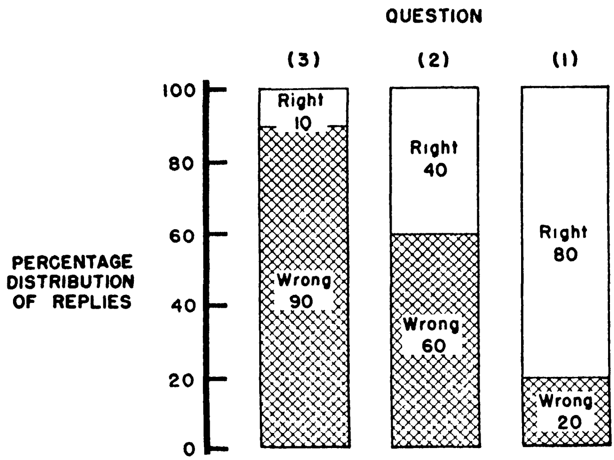


Figure 3

The bar chart representation. Another way of picturing the dichotomous scale of the sample of three items would be as follows: suppose that 80 per cent of the population got the first question right, 40 per cent got the second question right, and 10 per cent got the third question right. The univariate distributions of the three respective items could be shown by the bar charts in Figure 3.

The bars show the percentage distributions for the respective questions. The multivariate distribution for the three questions, *given that they form a scale for the population*, can also be indicated *on the same chart*, since all those who are included in the group getting a harder question right are also included in the group getting an easier question right. Thus, we could draw the bar chart over

again, but connect the bars with dashed lines in the fashion shown in Figure 4.

Here again we can see how the three questions are simple functions of the scores. From the marginal frequencies of the separate items, *together with the fact* that the items form a scale, we are enabled to deduce that 10 per cent of the people got a score of 3. The 10 per cent who got the hardest question right are included in those who got the easier questions right. This is indicated by the dashed line on the right, between the scores 2 and 3, which carries the same 10 per cent of the people (those with a score of 3) through the three

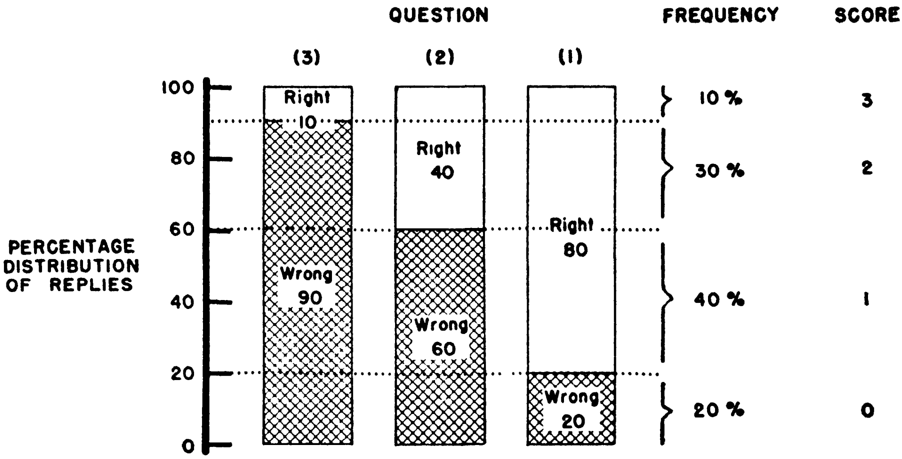


Figure 4

bars. The 40 per cent who got the second question right include the 10 per cent who got the hardest question right and 30 per cent out of those who got the hardest question wrong, but all 40 per cent got the easiest question right. This leaves us 30 per cent who got just the first and second questions right. And so on. Thus we can think of an ordering of the persons along a vertical continuum, and each dichotomy can be thought of as resulting in one additional *cut* on that continuum. All those above the cutting point get the question right, and all those below the cutting point get the question wrong. Thus, there is a one-to-one correspondence between the categories of an item and segments of the continuum. *Or we can say that each attribute is a simple function of the rank order along the continuum.*

If the "right" and "wrong" answer categories are separated, as in Figure 2, this bar chart representation assumes the pattern of a

parallelogram. Two basic steps are involved in this procedure, as will be described in detail in the next chapter. First, the questions are ranked in order of "difficulty" with the "hardest" question, i.e., the one that fewest persons got right, placed first and with the other questions following in decreasing order of "difficulty." Second, the people are ranked in order of "knowledge" with the "most informed" persons, i.e., those who got all questions right, placed first, the other individuals following in decreasing order of "knowledge." These two steps are the basic procedure for the scalogram board technique of scale analysis. The resulting parallelogram, assuming a scale is present, is shown in Figure 5.

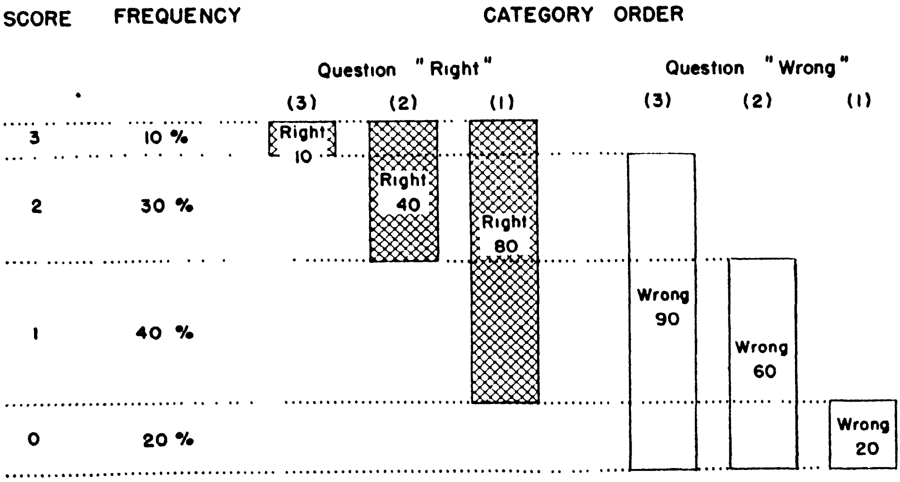


Figure 5

Zero-order correlations between items. It is because all the items in the sample can be expressed as simple functions of the same ordering of persons that they form a scale. Each item is perfectly correlated with or reproducible from the ordering along the continuum. However, the point correlations between the items are not at all perfect. For example, the fourfold table between the second and third items is as follows:

		Question 2 Right	*Question 2* Wrong	
Question 3	Right	10	0	10
	Wrong	30	60	90
		40	60	100

The point correlation between the two items is .41. As a matter of fact, the point correlation between two dichotomous items may be anything from practically zero to unity, and yet they may both be perfect simple functions of the same quantitative variable.[6]

An important feature of this fourfold table is the zero frequency on the upper right hand corner cell. Nobody who got the third question right got the second question wrong. Such a zero cell must always occur in a fourfold table between two dichotomous items which are simple functions of the same quantitative variable. This zero cell, furthermore, must occur in the column or row which contains the lowest frequency and in that cell which represents a "positive" answer on one question and a "negative" answer on the other question. Given only the marginal distributions of any two scale questions, it is possible to compute the frequencies in each cell of a correlation table of the two questions. In fact, *it is possible to construct the multivariate distribution of all the scale questions from a knowledge of their straight distributions alone.*[7] This can be done regardless of how many answer categories are retained for each question. An example of a scale using trichotomous items will be given later.

This requirement concerning zero-order correlation tables between items suggests a procedure for scale analysis based on these tables. However, zero-order relationships do not tell the whole story about the entire multivariate distribution. Actually, it is simpler to study the complete distribution by means of the scalogram board technique, or related techniques, than to study all the bivariate tables. It is helpful, however, to learn what a scale pattern signifies by exploring some of the consequences in terms of bivariate tables. While the zero-order correlation tables are not good to use as a practical technique for testing items for scalability, they are a good pedagogical device for understanding scales.

Zero cells for dichotomies. The presence of a zero cell in the proper place is a necessary but not sufficient condition for the existence of a

[6] A tetrachoric coefficient for the fourfold table above, assuming a bivariate normal distribution, would be unity. However, this is *not* the correlation between the items. It does not tell how well one can predict one item from the other. The tetrachoric coefficient expresses instead the correlation between two quantitative variables of which the items are functions, provided the assumptions of normality are true. The reason the tetrachoric is unity in this case is that the quantitative variables of which the items are functions are one and the same variable, namely, the scale variable. Notice, however, that the distribution of the scale variable according to the rank order is not at all normal. One of the contributions of scaling theory is to do away with untested and unnecessary hypotheses about normal distributions.

[7] The importance of this predictability for correlation analysis of attitude questions will be discussed later.

scale. If a correlation table of two attitude questions does *not* reveal this zero cell, then one can be certain that these two questions are not members of a single attitude continuum, i.e., do not have* but one meaning to the respondents. However, the existence of a zero cell in itself is insufficient proof of a single variable, even if the cell is in the proper position. First, the two questions must be judged to belong to the same content universe by virtue of their content. Second, a sample of two questions is too subject to sampling error (with respect to the universe of content) to provide an adequate test; the occurrence of the zero cells in their proper position should be found in the multivariate distribution of several questions. The desired pattern for more than two questions will become more clear to the reader as additional examples are given.

Let us translate this simple example into a problem of public opinion analysis. Suppose the issue to be studied is public opinion toward the continued maintenance of a large Army now that the war is won. Three questions dealing with this topic, together with hypothetical percentage distributions of replies, are given below.

1. In your opinion, is it necessary or unnecessary for the protection of the United States to have a strong Army?
   (a) It is necessary 70%
   (b) It is not necessary 30

2. Do you think the United States should or should not increase the present size of the Army?
   (a) It should 50%
   (b) It should not 50

3. If all other countries agree to disarm, do you think the United States should or should not maintain a large Army?
   (a) It should 20%
   (b) It should not 80

For the present, we will not deal with the problem of interpreting which of these percentages—20 per cent, 50 per cent or 70 per cent—represents the division of public opinion upon the issue of the size of the Army. This problem will be treated in detail in Chapter 7, "The Intensity Component in Attitude and Opinion Research."

Following the procedure outlined in relation to the statistical knowledge example given previously, we can test the hypothesis that these three questions form a scale by seeing whether the three questions are simple functions of the scale scores. One way of doing this would be to draw the bar chart diagram shown in Figure 5

and to compare the obtained frequency of the four scale types with the theoretical or expected frequency.

To draw this diagram of expected frequencies, we first place the questions in descending order from the question calling for the most extreme expression of "favorableness" toward a large Army, i.e., that question to which fewest persons reply "in favor of" a large Army, to the question calling for the least extreme expression of "favorableness" toward a large Army, i.e., that question to which the largest number of persons reply "in favor of" a large Army.[8]

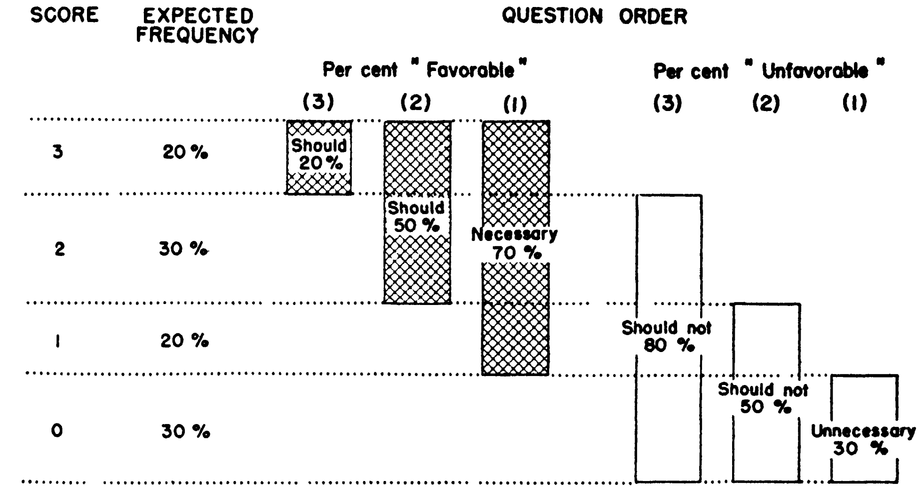


Figure 6

This ordering of the questions shows us immediately how many scale types should occur and what their expected frequencies should be, if we have a scale. Scoring each of the subjects and placing them in rank order of "favorableness" toward a large Army, i.e., the number of questions upon which they express opinions "in favor of" a large Army, should produce the parallelogram shown in Figure 6.

Thus, in order for these three questions to form a scale, all persons who felt that the United States *should* maintain a large Army even if all other countries disarm must also feel that the United States *should* increase the present size of the Army and that it was *neces-*

[8] The terms "favorable" and "unfavorable" are used in the sense of *more* or *less* favorable only. To divide such a continuum into "favorable" and "unfavorable" calls for the determination of a zero point—a problem which will be discussed in a later chapter.

sary for the protection of the United States to have a strong Army. Examination of the fourfold table between any two of the questions would have to reveal a zero cell in the corner representing the "favorable"-"unfavorable" cell of the lowest marginal frequency. For example, the frequencies of the cells in the cross tabulation between questions 2 and 3 are completely predictable from the marginals, as follows:

		Question 2 "Favorable"	*Question 2* "Unfavorable"	
Question 3	"Favorable"	20	0	20
	"Unfavorable"	30	50	80
		50	50	100

To conform to the scale pattern, there should be *no* respondents who feel that the United States *should* maintain a large Army even if all other countries disarm, but who think that the United States *should not* increase the present size of the Army. On the other hand, a respondent who feels that the United States *should not* maintain a large Army if all other countries disarm, may or may not favor an increase in the present size of the Army.

Given a scale pattern, the same predictability or reproducibility of intercorrelations between any questions is possible from a knowledge of the straight distributions. In addition, a knowledge of the scale score enables one to predict or reproduce the responses of any individual to each of the questions asked. The main condition for a scale is satisfied: a person with a higher rank than another is just as high or higher on each item.

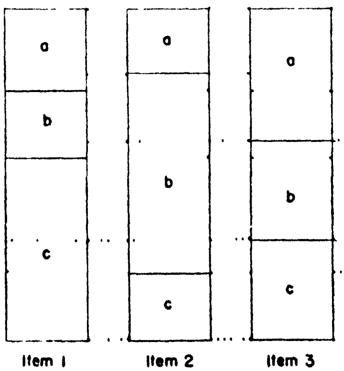
Example of a scale of trichotomies. The same technique of scale analysis outlined for dichotomous items applies equally well to items with any number of answer categories. In fact, this technique enables one to determine whether the different answer categories should be treated separately as representing meaningfully different replies or whether they should be combined. For example, in response to a question, "How important is it for the United States to have a large Army?" how should the answer categories "fairly important" and "not so important" be treated? Should they be left as separate categories, or should they be combined into a kind of neutral category, or should "fairly important" be combined with

"very important" while "not so important" is combined with "not at all important?" Scale analysis, using the technique outlined, will tell one quite automatically how these combinations should be made. This will be shown in several examples to be presented in a following chapter.

Let us see what a scale composed of trichotomous items would look like. Suppose three questions from the same attitude universe were asked, each of which had three answer categories, the answers to which distributed as follows:

	Question		
Answer categories	1	2	3
a	25%	20%	40%
b	20	60	30
c	55	20	30
	100%	100%	100%

The number of scale types for these three trichotomous items is seven, whereas the number of possible types for a nonscalable area is twenty-seven. This reduction in scale types is indicative of the highly restrictive nature of the scale pattern. (The number of scale types equals the *sum* of all item categories, less the number of questions, plus one, while the number of all possible types equals the *product* of all the item categories.) The characteristics of the scale types can be determined simply by joining the answer categories of the different questions, as below in Figure 7.



Scale type	Frequency	Scale score
aaa	20 %	6
aba	5 %	5
bba	15 %	4
bbb	5 %	3
cbb	25 %	2
cbc	10 %	1
ccc	20 %	0

Figure 7

Arranging the questions in order of the frequency of "favorable" responses (for all three answer categories) and arranging the subjects in order of scale scores (assume simple weights of two for *a*, one for *b*, and zero for *c*), the following bar chart parallelogram emerges for the ideal scale pattern (Figure 8).

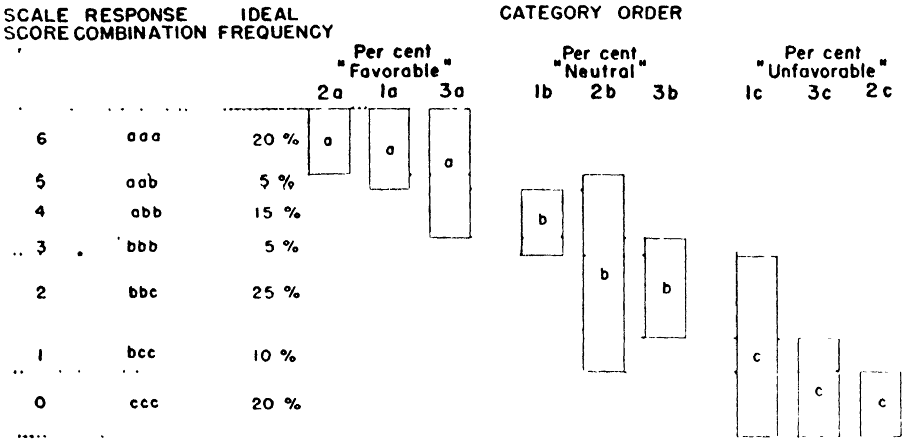


Figure 8

This pattern is completely predictable on the basis of the marginal distribution of answers to each of the questions. From a person's scale score it is possible to reproduce exactly how he answered each of the questions. From the marginal distribution of responses to any of the questions it is possible to construct any of the internal correlations between questions. The pattern of intercorrelation for questions 1 and 3, for example, would have to be as follows:

		Question 1 a	b	c	
Question 3	a	25	15	0	40
	b	0	5	25	30
	c	0	0	30	30
		25	20	55	100

The basic condition to be satisfied is that persons who answer a question "favorably" all have higher scale scores than persons who

answer the same question "unfavorably." This constitutes a rigorous definition of a scale. It provides a simple, objective technique for testing the existence of a single variable, that is, for determining whether the questions have the same meaning for all respondents.

The Measurement of Error[9]

The amount by which a scale deviates from the ideal scale pattern is measured by a *coefficient of reproducibility*. This coefficient is simply a measure of the relative degree with which the obtained multivariate distribution corresponds to the expected multivariate distribution of a perfect scale. It is secured by counting up the number of responses which would have been predicted wrongly for each person on the basis of his scale score, dividing these errors by the total number of responses and subtracting the resulting fraction from 1. As will be seen in the actual examples to be presented, the occurrence of errors is easily determined by visual inspection of the scale pattern. An acceptable approximation to a perfect scale has been arbitrarily set at 90 per cent reproducibility. Thus, if a scale consisted of five items tested on 100 people, the total number of responses would be $5 \times 100 = 500$. To secure a coefficient of reproducibility of at least .90, there could be at most 50 errors for the entire sample of 100 respondents on all five questions.

The coefficient of reproducibility of the universe can be observed only with sampling error if the scale pattern is not perfect. There may be error due to the sampling of items and error due to the sampling of people. The problem of item sampling is considered later. With respect to sampling deviations due to people, the following remarks may be made.

Reproducibility is computed by counting up the errors for each person on each item. If the sample consists of 100 persons and five items, the per cent reproducibility is based on 500 observations. If the errors of reproducibility are random, and if the population reproducibility is at least .90, then the standard error of a sample proportion of reproducibility is at most .013, which allows a deviation in the proportion of at most .040 at the three standard error level of confidence.

Such a calculation of a standard error is, of course, voided if the errors are not random, which may often be the case. Empirical experiments on samples of 100 cases each on five items, however,

[9] For a more detailed discussion of this problem see Chapter 8, "Problems of Reliability."

have consistently shown less than a variation of .04 when the combined sample reproducibility has been over .90. Occasionally an item has behaved a little differently for comparable samples of people. Raising requirements for reproducibility, and using items with more than two answer categories, reduces the possibility for such minor variations.

In some cases, a sample of 200 or more people may be necessary to get a clear picture of the situation.

Reproducibility is not a sufficient criterion for scalability. Reproducibility by itself is not a sufficient test of scalability. It is the principal test, but there are at least four other features that should be taken into account: (a) range of marginal distributions, (b) pattern of errors, (c) number of items in the scale, (d) number of response categories in each item.

(a) *Range of marginal distributions:* The reproducibility of any individual item can never be less than the percentage of respondents falling into a single answer category of that item, regardless of whether or not a scale exists. For example, if a dichotomous item has 80 per cent of the people in one category and 20 per cent in the other, there cannot be less than 80 per cent reproducibility in reproducing that item from a rank order obtained from all the items, regardless of the scalability of the set of items as a whole. Thus, if a sample comprises only items with extreme kinds of dichotomizations, reproducibility will be automatically high for that sample, regardless of the scalability of the universe. Therefore, to test a universe for scalability, attempts should be made to include in the sample as wide a range of marginal distributions as possible, and specifically to attempt to include items with marginals around 50–50.

(b) *Pattern of errors:* If an area is scalable with but 10 per cent error (and not artificially so because of extreme marginals), this implies that there is but one dominant variable in the area along which to order the persons. The errors of reproducibility may be caused either (a) by one or two other variables of lesser magnitude that may be in the area, or (b) by many small variables.

The existence of one or two additional small variables as opposed to many small variables in the area is indicated by *nonscale types* in the scale pattern which occur with sufficient frequency to be noticed, but not with enough frequency to impair substantially the reproducibility of the area from only the dominant variable. If such definite nonscale types exist, then the multiple correlation of an outside variable with the whole area would *not* be quite equiv-

alent to the simple correlation with rank order on the dominant variable, and would be attained only by taking the nonscale types into account.

On the other hand, if error of reproducibility is random, then the multiple correlation of any outside variable on the area will be precisely equal to the simple correlation with the rank order on the area. This property, it is important to note, holds *no matter how low the reproducibility is.* Some areas which are not scalable are called *quasi scales*; their reproducibility may not be high but their errors occur in a sort of gradient. This means that although they lack the essential property of a scale—rank order cannot reproduce persons' characteristics on the items in the area very well—nevertheless the rank order is perfectly efficient for relating any outside variable to the area.

The difference between random errors, nonscale errors, and gradient errors will be discussed in Chapter 5, "The Utility of Scalogram Analysis."

(c) *Number of items:* The more items included in a scale, the greater is the assurance that the entire universe of which these items are a sample is scalable. If the items are dichotomous (or dichotomized from more than two categories as a result of the scale analysis), it is probably desirable that at least ten items be used, with perhaps a lesser number being satisfactory if the marginal frequencies of several items are in the range of 30 per cent to 70 per cent. Just four or five items, with marginal frequencies outside such a range, would not give much assurance as to how scalable the universe was, no matter how scalable the sample might be. In practice, ten or more items can be used on a pretest to determine whether or not a universe is scalable but fewer items can be used in the larger study—if the universe is shown to be scalable by the pretest—to obtain the number of ranks necessary for the amount of discrimination between people required by the study.

(d) *Number of response categories:* The more response categories for items included in a scale, the greater is the assurance that the entire universe is scalable. A caution to be observed in combining response categories to reduce error is to make sure that the reduction in error is not just a consequence of obtaining new extreme marginal frequencies (e.g., 90–10) that do not permit much error (see *a* above). Equally important is the fact that keeping answer categories separate, while it will usually increase the amount of error, decreases the possibility of a scale pattern appearing if in fact

the universe is nonscalable. For example, four dichotomous items with high reproducibility do not provide as dependable an inference concerning the scalability of an area as would four trichotomous items which were equally as reproducible. It is especially important to keep as many response categories as possible when the total number of items is small. The more categories that can remain uncombined, the more reliable is the inference that the universe from which they come is scalable.

In many cases, of course, sufficient reproducibility may exist without combining categories. In such a case categories may be combined anyhow for convenience in final scoring. Combining categories in such a case does not disturb rank order except that two adjacent ranks are merged.

The Universe of Attributes[10]

A basic concept of the theory of scales is that of the universe of attributes. In social research, a universe is usually a large class of behavior such as described in the introduction. The universe is the concept whose scalability is being investigated, like marital adjustment, opinion of British fighting ability, knowledge of arithmetic, etc. The universe consists of all the attributes that define the concept. Another way of describing the universe is to say it consists of all the attributes of interest to the investigation which have a common content, so that they are classified under a single heading which indicates that content.

An important consideration of the present theory of scales becomes that of the sampling of items. In studying any attitude or opinion, there is an unlimited number of questions or question wordings which could be used. Any question asked in an attitude or opinion survey is ordinarily but a single sample of indefinitely many ways the question could be put. It is well known that changing the wording of the questions, changing the order of presentation of questions, changing order of check lists of answers, etc., can yield apparently different results in the responses. Conceivably, one could ask questions which would secure "favorable" replies ranging from 0 to 100 per cent, depending upon the extremeness of the statement that the respondents are asked to approve or disapprove. It is,

[10] The words *population* and *universe* are ordinarily used interchangeably in statistical literature. For scales, it is necessary to refer both to a complete set of objects and to a complete set of attributes, so it will be convenient to reserve *population* for the former, and *universe* for the latter. In social research, the objects are usually people, so that *population* is appropriate for them.

therefore, essential to inquire into the nature of the *universe of all possible questions of the same content*, and to determine what inferences can be made about that universe that will not depend on the particular sample of questions used.

Scalogram theory shows that if the universe contains but a single variable, that is, if all questions have but a single content ordering, then the same rank order of the individuals upon this content will be obtained regardless of which sample of questions is selected from the universe. The problem of sampling of items thus has a simple solution for the case of a scalable universe.

An important property of a scalable universe is that the ordering of persons based on a sample of items will be essentially the same as that based on the universe. If the universe is scalable, the addition of further items merely breaks up each type given by the sample into more differentiated types. But it would not interchange the order of the types already in the sample. For example, in Figure 8 above, type 5 would always have a higher rank order than type 4. People in type 5 might be ordered within the type into more subcategories; people within type 4 might be ordered into more subcategories; but all subcategories within 5 would remain of higher rank than all those in type 4. This may be seen in reverse, for example, by deleting one of the questions or by combining answer categories so as to make a trichotomous question dichotomous, and noticing that all that is accomplished is to collapse the number of types to a smaller number so that two neighboring types may now become indistinguishable; but any types two steps apart would still remain in the same order with respect to each other.

Hence, we are assured that if a person ranks higher than another person in a sample of items, he will rank higher in the universe of items. This is an important property of scales, that *from a sample of attributes we can draw inferences about the universe of attributes.*

One of the criteria for selecting a sample of items is to choose a sample with enough categories to provide a desired amount of differentiation between individuals.[11] Thus if individuals are desired to be differentiated, say, into only ten groups, items should be chosen which will yield ten types.[12] The shape of the distribution

[11] The number of possible scale types may be determined as follows: add unity to the total number of categories in all questions (after combination) and subtract the number of questions.

[12] We are of course not considering problems of reliability in the sense of repeated observations of the same attributes. For convenience, we are tacitly assuming perfect reliability. See Chapter 8, "Problems of Reliability."

of the rank orders in a sample of items will of course depend upon the marginal frequencies of the items selected. One sample of items may yield a distribution of one shape; another sample may yield a completely different shape. This need not be a matter of concern, since our primary interest lies in the ordering of people, not the relative frequency of each position.

It might be asked how one can know the universe forms a scale if all one knows is a sample from the universe.

At present it seems quite clear that in general the probability of finding a sample of items to form a scale by chance for a sample of individuals is quite negligible, even if there are as few as six dichotomous items in the sample and as few as one hundred individuals.[13] It seems quite safe to infer in general that if a sample of items is selected without knowledge of their empirical interrelationships and is found to form a scale for any sizable random sample of individuals, then the universe from which the items are selected is scalable for the entire population of individuals. This problem has already been discussed under "Measurement of Error."

The relativity of scales. A universe may form a scale for a population at a given time and may not at a later time. Such a change in time would tend to indicate that the change is one of *kind*, rather than *degree*. A new meaning has been added to the previous single variable. For example, the items in a scale of expression of desire of American soldiers to go back to school after the war may not prove to be scalable if they were asked once more at the close of the war.

Conversely, a universe may not be scalable at one time, but scalable at another. This would indicate a change in the structuralization of the attitude from many dimensions to one dimension, or that an "unstructured" attitude has become "structured."

A universe may form a scale for one population of individuals, but not for another. Or, the items may form scales for two populations in different manners. For example, a sample of items of satisfac-

[13] To work out the complete probability theory would require two things: first, a definition of a sampling process for selecting items, and, second, a definition of what is meant by a scale not existing. A definition of the sampling process is difficult because items are ordinarily developed intuitively. Stating a null hypothesis that a scale does not exist leads to many possible analytical formulations, for different limiting conditions may be imposed upon the multivariate distribution of the items. For example, should the marginal frequencies be considered fixed in all samples, should the bivariate frequencies be considered fixed, etc.? These are questions which may become clearer as the theory of scaling develops, and in return may clarify our conceptions of what observation of social phenomena implies.

tion with Army life which formed a scale for combat outfits in the Air Force did not form a scale for men in the technical schools of the Air Force. The structure of camp life for these two groups was too different for the same items to have the same meaning in both situations.

A universe may not form a scale for the total population, but still form a scale for subgroups of that population. The essential definition of a scale is that of "single-meaning," and while a series of questions may contain different meanings to a cross section of the population, they may contain only a single meaning for some subgroup of that population. However if a scale is obtained for a cross section of the population, then that same scale pattern necessarily holds for all major subgroups.

If a universe is scalable for one population but not for another population, we cannot compare the two populations in degree and say that one is higher or lower on the average than another with respect to the universe. They differ in more than one dimension, or in kind of attitude rather than in degree of "favorableness" on the same attitude. It is only if two groups or two individuals fall into the same scale that they can be ordered from higher to lower. A similar consideration holds for comparisons in time. An important contribution of the present theory of scaling is to bring out this emphasis quite sharply.

Content alone defines the universe. Before the structure of a universe can be analyzed, the universe itself must be defined. Let us take an example from opinion research where it is desired to observe the population of individuals in a standardized manner by a check list of questions. The behavior of interest to the investigation is responses of individuals to such questions. Suppose the universe of items consists of all possible questions which could be asked in such a list concerning the fighting ability of the British. Such questions might be: "Do you think the British Army is as tough as the German Army?" "Do you think the RAF is superior to the Luftwaffe?" etc. (We do not pause here for problems of wording, interpretation, and the like. The reader is urged rather to focus on the general outline we are trying to establish.) There may be an indefinitely large number of such questions which belong in the universe; and in a particular investigation, ordinarily only a sample of the universe is used.

An attribute belongs to the universe by virtue of its content. The investigator indicates the content of interest by the title he

chooses for the universe, and all attributes with that content belong in the universe. There will, of course, arise borderline cases in practice where it will be hard to decide whether or not an item belongs in the universe. The evaluation of the content thus far remains a matter that may be decided by consensus of judges or by some other means. It may well be that the formal analysis for scalability may help clarify uncertain areas of content. However, we have found it most useful at present to utilize informal experience and consensus to the fullest extent in defining the universe.

It should be pointed out that an area of qualitative data which has been carefully thought through and judged to comprise a homogeneous universe of content does not necessarily form a scale. The concepts of universe and of scale are distinct and separate. If a universe is not a scale, it cannot be represented by a single rank order. In some cases, scale analysis may suggest that there are two or more subareas in the universe which might be scalable separately. Then the universe could be represented by *several* scale variables, by giving each person a rank order on each of the scalable subareas. It may happen that a sample of items is analyzed for a group of people and is not found to be scalable; but one or more subsets of the items seem to be scalable separately. Finding scalable subsets of items may sometimes imply that the original universe of content can be divided into subuniverses, at least one—or perhaps all—of which are scalable separately. To test the hypothesis that a scalable subset is part of a scalable subuniverse, it is necessary to show that the content of this subuniverse is ascertainable by inspection, and is distinguishable by inspection from that of the rest of the universe. A practical procedure to test this hypothesis might be as follows: construct new items of two types of content, one type which should belong in the original universe but should not belong to the scalable subuniverse. If the new items designed for the apparently scalable subuniverse do scale, and scale together with the old subset; and if the new items designed not to be in this subuniverse do not scale with the subscale; then the hypothesis is sustained that a subuniverse has been defined and has been found scalable. Each hypothesized subuniverse might be tested this way.

It often happens, when about a dozen questions in an area are pretested, that all but one or two are found to be scalable. There are at least two possibilities as to why some items do not scale while others in the same area do: (a) the universe is not scalable as a whole,

but contains a scalable subuniverse; (b) the items have been imperfectly constructed. This latter reason is so easy and glib that it is best to avoid it as much as possible. If the vast majority of a sample of items do scale, then it may be plausible to blame faulty construction for the nonscalability of one or two items. This hypothesis can be tested by rebuilding the apparently faulty items and retesting them. In practice, if the vast majority of a sample of items do scale, the one or two items that do not may often be ignored. If their construction is faulty, there are sufficient items without them to establish rank order for the people; and if they really represent a nonscalable part of the original universe, this part may be so small as not to be essential to the study.

Scale analysis does not define content. The question may be asked: to what extent, if any, can scale analysis be used to arbitrate differences of opinion with respect to content, that is, as to whether or not a given item belongs in an area? The answer is quite simple. Scale analysis as such gives no judgment on content;[14] *it presumes that the universe of content is already defined*, and merely tests whether or not the area is representable by a single variable. It might serve as an *auxiliary* argument with respect to content in the special case where there is controversy over but one or two items of a large sample of items in which the remaining items are scalable. In such a case, if the items in doubt do scale with the others, this may be taken as additional evidence supporting the contention that they belong in the area. It should be emphasized that this kind of inference is but *auxiliary*—there must be a cogent initial argument based on content if an item is to be classified in an area. Sheer scalability is not sufficient; an item may happen to scale with an area, and yet not have the content defining the area—it may be a correlate rather than part of the definition.

An important emphasis of our present approach is that a criterion for an attribute to belong in the universe is *not* the magnitude of the correlations of that item with other attributes known to belong in the universe. Attributes of the same type of content may have any size of intercorrelations, varying from practically zero to unity.[15]

[14] More generally, *no* correlation analysis as such determines content; it studies only formal relationships. If x relates perfectly with y, that does not mean x is identical with y. If it is known that the correlation between x and y is .6, that alone does not help to name the content of either x or y.

[15] That correlations are no criterion for content has been quite well known. See, for example, R. F. Sletto, *Construction of Personality Scales by the Criterion of Internal Consistency* (Sociological Press, Hanover, N.H., 1937).

Methods of observation. Let us assume that somehow we have a universe of attributes and a population of individuals defined. Next, observations are made as to the behavior of the population with respect to the universe. (In practice this will often be done only with samples. A sample of individuals from the population will have their behavior observed on a sample of attributes from the universe.) How the observations are to be made is of no theoretical concern here. In opinion research and other fields, questionnaires and schedules have been used. But any technique of observation which yields the data of interest to the investigation may be used. Such techniques for the social and psychological sciences might be case histories, interviews, introspection, and any other technique from which observations may be recorded. The important thing is not how the observations were obtained, but that the observations be of central interest to the investigation.

Use of a questionnaire implies that the investigator is interested in a certain type of universe of verbal behavior. Participant observation may imply that the investigator is interested in a certain type of nonverbal behavior. Such distinct universes may each be investigated separately. It may often be of interest to see how well one universe correlates with another, but such a correlation cannot be investigated until each universe is defined and observed in its own right.

A good deal of attention has been given by various research workers to the types of observations recorded. Thurstone has approached the problem of attitude measurement by asking essentially, "What are your opinions?" [16] Other research workers have attempted to get at attitudes by asking different questions. For example, Bogardus asks, "What social relations are you willing to tolerate with these people?",[17] while Pace asks, "What would you do under these circumstances?" [18] Ford studied an attitude area by asking about past behavior and concludes, "One may say that the scales for estimating . . . experiences are really attitude scales in disguise." [19] And so in the literature of attitude analysis

[16] L. L. Thurstone and E. J. Chave, *The Measurement of Attitude* (University of Chicago Press, Chicago, 1929).

[17] See an application of this in Stuart C. Dodd, "A Social Distance Test in the Near East," *American Journal of Sociology*, Vol. 41, No. 2 (September 1935), p. 195.

[18] C. R. Pace, "A Situations Test to Measure Social-Political-Economic Attitudes," *Journal of Social Psychology*, Vol. 10, No. 3 (August 1939), pp. 331–344.

[19] R. N. Ford, "Scaling Experience by a Multiple-Response Technique: A Study of White-Negro Contacts," *American Sociological Review*, Vol. 6, No. 1 (February 1941), p. 21.

we find much space devoted to the construction of areas by means of different types of items and check lists. All of these different question forms represent only differences in techniques of observing the attitude universe. There is no a priori reason to assume that if the attitude universe is, say, attitude toward the British, any one way of securing observations of this universe is better than any other. In fact, if the universe is scalable, all types of questions may be used indiscriminately with the same result as far as the rank order of individuals in that universe is concerned. If, however, one series of items (for example, questions about past experiences with the British) does not scale with another series of items concerning the British (for example, behavior in hypothetical situations), then these two series may be regarded as two scalable subuniverses of a larger but nonscalable universe.

Observations made in public opinion polls. It is with regard to the manner of observation that public opinion polls have, perhaps, been most often differentiated from attitude scales. A question concerning public opinion, it is argued, is more specific and controversial than an attitude question. This specificity has led public opinion pollsters to place a great deal of stress upon the "unbiased" construction of a single meaningful question. Since the validity of the poll question depends upon its manifest content, a great deal of concern is shown over the "meaning" of the single question. It is felt that practically every word must be tested for understandability, bias, single meaning, etc., while check-list answer categories must be presented in reverse form or combined in different ways, etc. Actually, it is this susceptibility of the single poll question to slight changes in question wording that reinforces the position of the present approach that *attitude questions and opinion questions represent simply different forms of observation.*

Any single opinion question is only a sample of one from the whole universe of questions in the opinion area. With regard to differences in "specificity" between attitudes and opinions, it is our position that any area of behavior can usually be regarded as a subuniverse of some larger area, and can itself often be divided further into subareas. In other words, there is no means of determining when an attitude is specific enough to be called an opinion. Our definition of both involves the determination of verbal or nonverbal behavior in some area, which may be relatively specific or relatively general. In either case, of attitudes or of opinions, validity will depend upon the adequate sampling of the entire universe of ques-

tions in the area of interest. Scale analysis provides an objective test of whether or not any particular opinion poll question contains but a single dimension of meaning that is common to all similar questions that could have been asked in its place.

The examples of scales to be given later happen to comprise observations made by means of questionnaires. It should not be inferred, however, that scaling refers only to that technique. *Scale analysis is a formal analysis, and hence applies to any universe of qualitative data of any science, obtained by any manner of observation.*

Summary

1. The need for scale analysis arises out of the fundamental problem of attitude scaling and opinion polling of how to determine the dimensions of meaning which the questions asked have for the respondents. Scale analysis affords a rigorous test for the existence of single-meaning for an area and provides a rank order of individuals for such areas as are found scalable.

2. Basic to the present scale theory is the concept of sampling the attitude or opinion universe. An unlimited number of questions could be asked in any area; the problem is one of selecting a sample of questions which are representative of all possible questions that might have been asked. Scale analysis affords a test of this sampling.

3. Scale analysis applies equally well to the study of attitudes as to the study of opinions. More generally, any technique of observation—questionnaires, interviews, participant observation, etc.—of any form of verbal or nonverbal behavior which is qualitative, yields data which may be studied by scale analysis.

4. Scale analysis tests the hypothesis that a group of people can be arranged in an internally meaningful rank order with respect to an area of qualitative data. A rank order of people is meaningful if, from the person's rank order, one knows precisely his responses to each of the questions or acts included in the scale.

5. More precisely, the multivariate frequency distribution of a universe of attributes or items for a population of objects or people is a scale if it is possible to derive from the distribution a quantitative variable with which to characterize the objects such that each attribute in the universe is a simple function of that quantitative variable.

6. The scalogram board technique for determining the existence of a scale involves two basic steps: (1) the questions and answer

categories are ranked in a preliminary order of extremeness with the "most extreme" category, i.e., the one which is endorsed by fewest people, placed first and the other categories following in decreasing order of "extremeness," and (2) the people are ranked in order of "favorableness" with the "most favorable" persons, i.e., those who answer all questions "favorably," placed first and the other individuals following in decreasing order of "favorableness." The resulting pattern, if a scale is present, will be a parallelogram.

7. There is an unambiguous meaning to the order of scale scores and the order of categories within each item. An object with a higher score than another object is characterized by higher, or at least equivalent, values on each attribute. Similarly, one category of an item is higher than another if it characterizes persons all of whom are higher on the scale.

8. From the multivariate distribution of a sample of attributes for a sample of objects, inferences can be drawn concerning the complete distribution of the universe for the population.

a. The hypothesis that the complete distribution is scalable can be adequately tested with a sample distribution.

b. The rank order among objects according to a sample scale is essentially that in the complete scale.

9. Perfect scales are not found in practice.

a. The degree of approximation to perfection is measured by a *coefficient of reproducibility*, which is the empirical relative frequency with which values of the attributes do correspond to intervals of a scale variable.

b. In practice, 90 per cent perfect scales or better have been used as efficient approximations to perfect scales.

10. The predictability of any outside variable from the scale scores is the same as the predictability from the multivariate distribution with the attributes. The zero-order correlation with the scale score is equivalent to the multiple correlation with the universe. Hence, *scale scores provide an invariant quantification of the attributes for predicting any outside variable whatsoever.*

11. Scales are relative to time and to populations.

a. For a given population of objects, a universe may be scalable at one time but not at another.

b. A universe may be scalable for one population but not for another, or it may not be scalable for an entire population, but scalable for a subpopulation. However, if a universe is scalable for an entire population, it will be scalable for all major subpopulations.

c. Comparisons with respect to degree can be made only if the same scaling obtains in both cases being compared.

Bibliography on Scalogram Analysis

1. Clark, K. E., and Kriedt, P. H., "An Application of Guttman's New Scaling Techniques to an Attitude Questionnaire," *Educational and Psychological Measurement*, Vol. 8, No. 2 (Summer 1948), pp. 215–223.
2. Edwards, A. L., and Kilpatrick, F. P., "Scale Analysis and the Measurement of Social Attitudes," *Psychometrika*, Vol. 13, No. 2 (June 1948), pp. 99–114.
3. Edwards, A. L., and Kilpatrick, F. P., "A Technique for the Construction of Attitude Scales," *Journal of Applied Psychology*, Vol. 32, No. 4 (August 1948), pp. 374–384.
4. Festinger, L., "The Treatment of Qualitative Data by 'Scale Analysis,'" *Psychological Bulletin*, Vol. 44, No. 2 (March 1947), pp. 149–161.
5. Goodenough, W. H., "A Technique for Scale Analysis," *Educational and Psychological Measurement*, Vol. 4, No. 3 (1944), pp. 179–190.
6. Guttman, L., "The Quantification of a Class of Attributes: A Theory and Method of Scale Construction," in P. Horst et al., *The Prediction of Personal Adjustment* (Social Science Research Council, New York, 1941), pp. 319–348.
7. Guttman, L., "A Basis for Scaling Qualitative Data," *American Sociological Review*, Vol. 9, No. 2 (April 1944), pp. 139–150.
8. Guttman, L., "The Cornell Technique for Scale and Intensity Analysis," *Educational and Psychological Measurement*, Vol. 7, No. 2 (Summer 1947), pp. 247–280.
9. Guttman, L., "Suggestions for Further Research in Scale and Intensity Analysis of Attitudes and Opinions," *International Journal of Opinion and Attitude Research*, Vol. 1, No. 1 (March 1947), pp. 30–35.
10. Guttman, L., and Suchman, E. A., "Intensity and a Zero Point for Attitude Analysis," *American Sociological Review*, Vol. 12, No. 1 (February 1947), pp. 57–67.
11. Guttman, L., "On Festinger's Evaluation of Scale Analysis," *Psychological Bulletin*, Vol. 44, No. 5 (September 1947), pp. 451–465.
12. Noland, E. W., "Worker Attitude and Industrial Absenteeism: A Statistical Appraisal," *American Sociological Review*, Vol. 10, No. 4 (August 1945), pp. 503–510.
13. Shapiro, G., "Myrdal's Definitions of the 'South': A Methodological Note," *American Sociological Review*, Vol. 13, No. 5 (October 1948), pp. 619–621.
14. Suchman, E. A., and Guttman, L., "A Solution to the Problem of Question 'Bias,'" *Public Opinion Quarterly*, Vol. 11, No. 3 (Fall 1947), pp. 445–455.

CHAPTER 4

THE SCALOGRAM BOARD TECHNIQUE FOR SCALE ANALYSIS[1]

FROM the previous examples of ideal scales, it is apparent that techniques to determine whether or not the basic parellelogram pattern of responses is present in a given set of data can be very simple. The multivariate frequency distribution of responses should be such that from the scale score one can tell exactly what responses were made to each of the questions in the scale. This will be the case where the responses to each question can be expressed as a simple function of the scale scores. To determine whether or not a series of questions forms a scale, it is only necessary to test to what extent question responses are reproducible from scale scores.

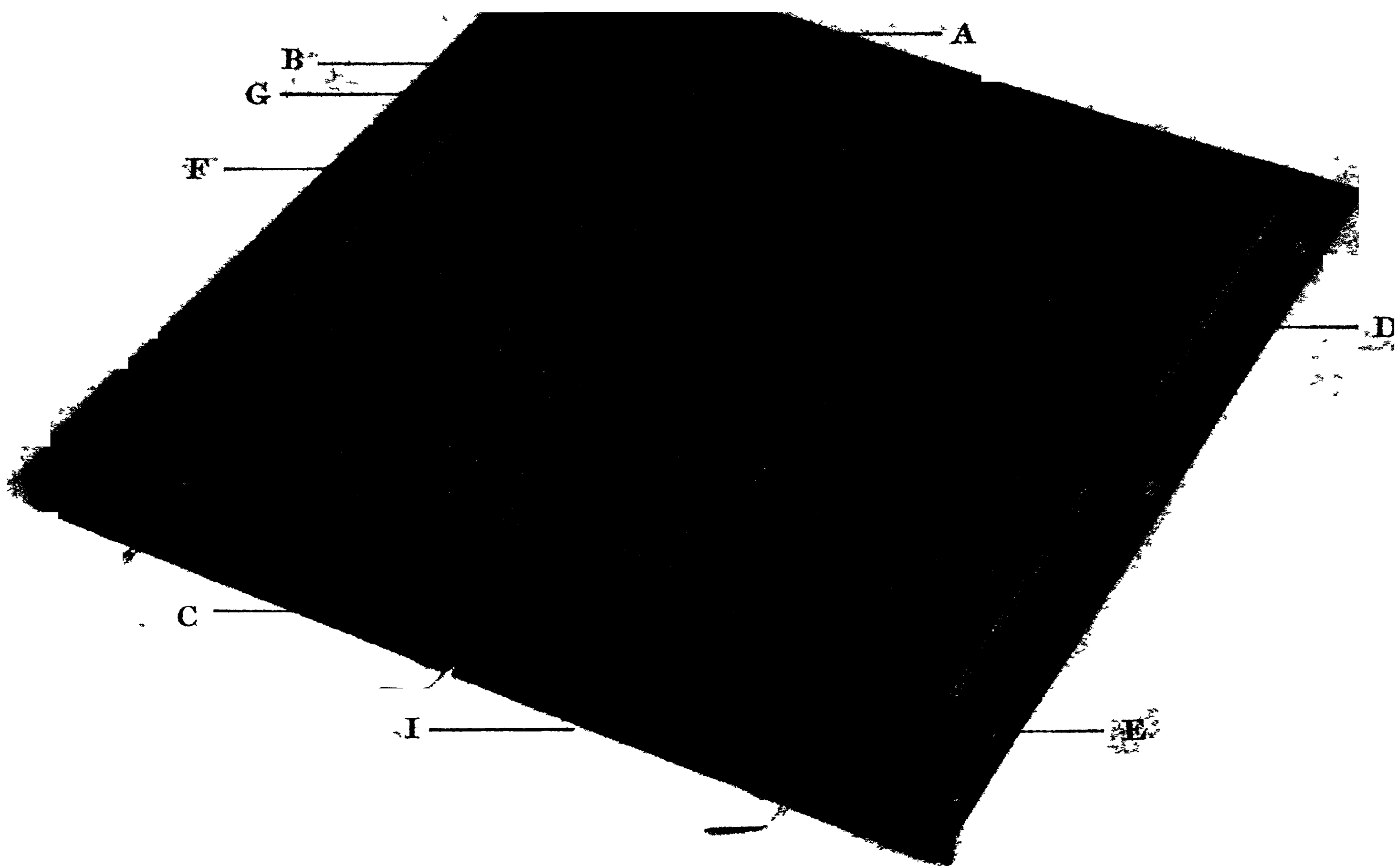
This may be done in several different ways. Four specific techniques which have been used successfully are: (1) the scalogram board technique, (2) the Cornell trial-scoring and graphic technique,[2] (3) the tabulation technique,[3] and (4) the least squares method.[4] All four techniques produce essentially the same results, since they have the same basic theory. The last three methods have been presented elsewhere and will not be discussed here. The first method, the scalogram board technique, has received the most use to date, being the technique employed by the Research Branch. This is the method that will be described in detail in the present chapter.

[1] By Edward A. Suchman.

[2] Louis Guttman, "The Cornell Technique for Scale and Intensity Analysis," *Educational and Psychological Measurement*, Vol. 7, No. 2 (Summer 1947), pp. 247–280.

[3] Ward H. Goodenough, "A Technique for Scale Analysis," *Educational and Psychological Measurement*, Vol. 4, No. 3 (1944), pp. 179–190.

[4] Louis Guttman, "The Quantification of a Class of Attributes: A Theory and Method of Scale Construction," in P. Horst et al., *The Prediction of Personal Adjustment* (Social Science Research Council, New York, 1941), pp. 319–348.


A
B
G
F
D
C
I
E

The Scalogram Board[5]

The scalogram board is simply a device which permits the shifting of the rank order of both respondents and question categories. By arranging answer categories in ascending order of frequency of "favorable" responses and by arranging respondents in descending order of "favorableness," it is possible to determine *visually* whether the required parallelogram pattern exists. This ability to shift item order and respondent order at will also permits the use of the scalogram board for many purposes other than scale analysis. Many problems involving an analysis of multivariate frequency distributions of qualitative data can be successfully handled visually by means of the scalogram boards.

Two scalogram boards are used, in order to provide both vertical (for people) and horizontal (for response category) mobility. The two boards are identical in construction. Each board serves as a base upon which 100 removable wooden slats are held in place by a frame attached to the base. When the frame is off, these strips of wood can be removed and shifted up or down the length of the board. Each strip has 100 holes bored equidistant and sufficiently deep to hold an 118 calibre shot which has been silver-coated. The shots are used to indicate each individual's response to each question. The board and slats are stained black so that the contrast of the silver balls against the black board enables one to photostat each piece of work. Since each board consists of 100 movable wooden strips, each strip containing 100 holes, the board is equipped to handle the responses of 100 individuals to items with a maximum total of 100 categories.

Legend for Figures 1, 2, and 3

A, B, C —— Frames attached to base piece (hold slats in place)
D —— Removable frame (When this frame is off, slats can be taken out and moved up or down.)
E —— Fill-in strip (used to tighten slats and align them for transfer to a second board)
F —— Alignment strip (used to align boards 1 and 2 for transfer)
G —— Movable slats—numbered 1 to 100
H —— Screw and nut (used to tighten and align slats)
I —— Base

[5] The scalogram board was designed by Louis Guttman.

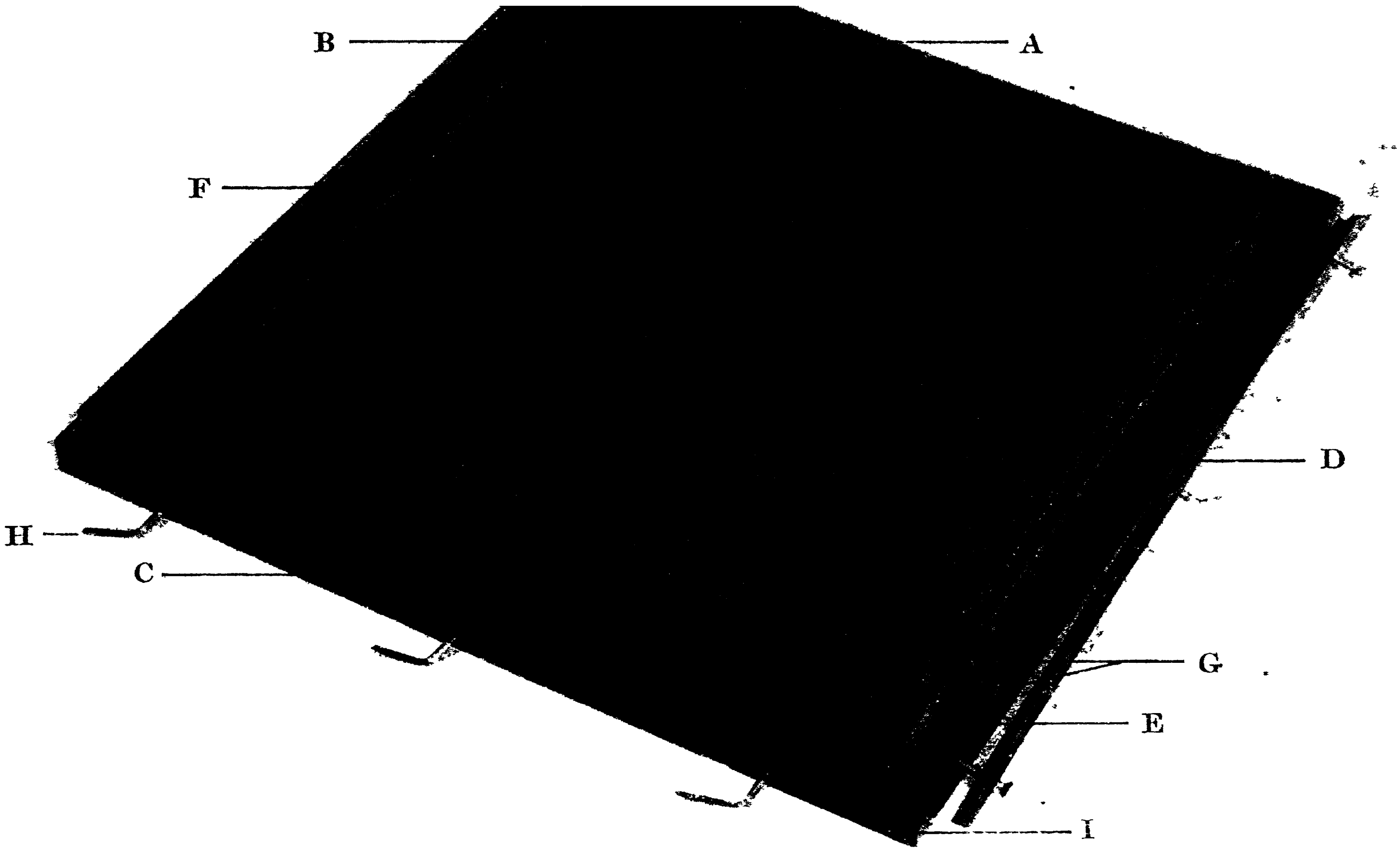


Figure 2 Scalogram board (showing slats ready to be shifted).

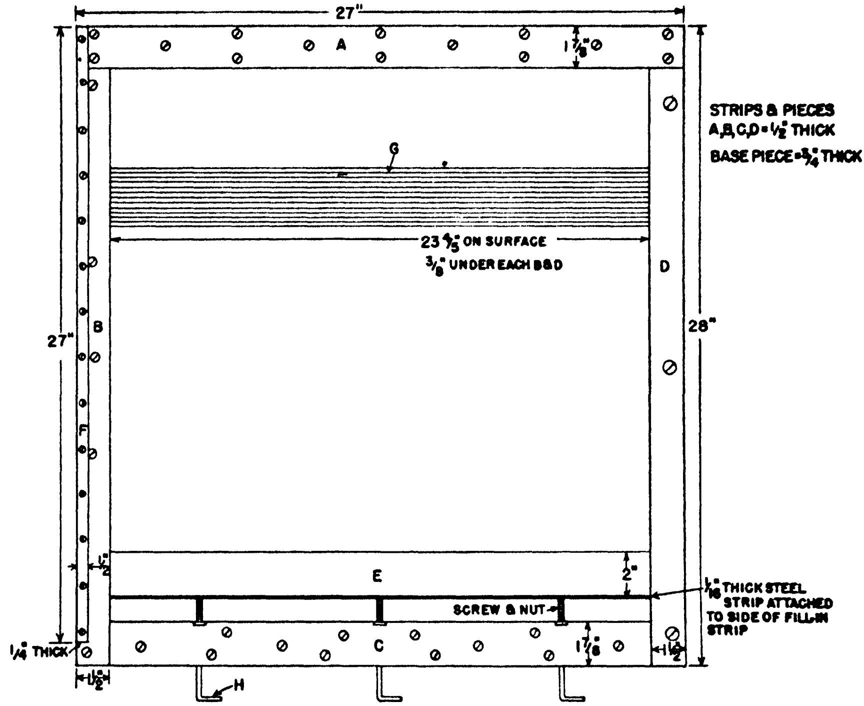


Figure 3. Diagram for construction of scalogram board

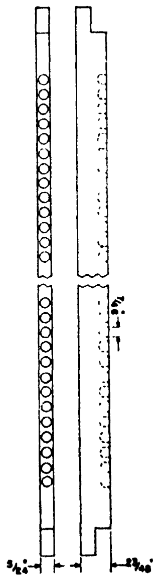


Figure 4. Diagram for construction of removable slats.

Pictures of the board are shown in Figures 1 and 2, while a diagram for the construction of the board is given in Figures 3 and 4. Although these are not shown in the pictures, it has been found helpful to attach strips numbered from 1 to 100 across the top of the board (Frame A) and down the side of the board (Frame B) to show the rank order of respondents and items.

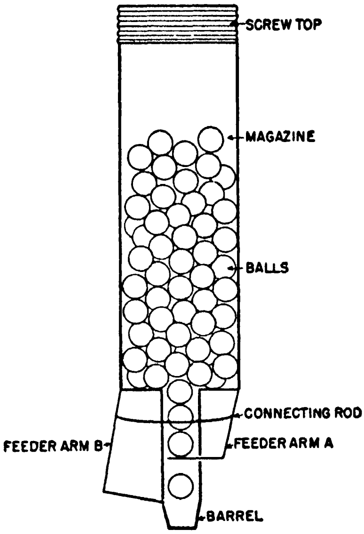


Figure 5. Dropper for insertion of balls.
How the Dropper Works: The magazine is kept filled with balls. As Feeder Arm A is pressed in, Feeder Arm B is pushed back and a ball is released. Feeder Arm A in the meantime prevents other balls in the barrel and magazine from being released. Relaxing pressure on Feeder Arm A permits it to spring back into place, closing the barrel opening by means of Feeder Arm B, and permitting another ball to fall between Feeder Arms A and B ready for release when Feeder Arm A is pressed in again.

The balls are inserted into the holes by means of a dropper which expels one ball at a time. The design of such a dropper[6] is shown in Figure 5. To remove the balls from the holes, it is only necessary to pass over them with a magnet.

The space between rows of holes is the same as between columns of holes on each board. This is so the holes of a column of one board will fit precisely over the holes of a row of the other board. Balls can be transferred from the holes of one board into the holes of the second board simply by putting the empty board face down on the board containing the balls, holding both boards together,

[6] The dropper was designed and constructed by Fred Sheffield.

and turning them over. The balls will then fall from the holes in Board 1 into the corresponding holes in Board 2. The balls are always in only one of the boards.

In setting up a scaling problem, the balls are first put into one board to arrange the individuals in rank order. The second board is used afterwards to arrange the categories in rank order. The removable slats of the first board permit the shifting of individuals, at the same time holding the order of question categories constant, while the removable slats of the second board permit the shifting of categories with the rank order of individuals remaining constant.

In order to change from shifting individuals to shifting categories, it is necessary to superimpose the two boards so that the slats of the two boards run at *right angles* to each other, and to turn the boards over so that the balls fall out of the holes of one board into the holes of the second board. For example, if Board 1 is used to arrange individuals, each removable slat represents the responses of one single individual to all the questions. These slats, let us say, run horizontally, that is, they can be moved up or down on the board. (See Figure 6.) The order in which the questions and answer categories appear is fixed, and the individual respondents can be moved up or down without disturbing the question order.

To change the category order it is necessary to transfer the balls to Board 2. This is done by placing Board 2 over Board 1, with the slats running at right angles. The two boards are then held together and turned over. In this way the balls are transferred from Board 1 to Board 2. (See Figure 7.) Again, the presence of a ball represents the answer of a respondent to a question. But on this board, the movable slats run in the opposite direction and each slat represents a question-answer category. The rank order of respondents is fixed, and the order of the categories can now be changed without disturbing the respondent order. After rearranging category order, to change back to Board 1 in order to study the effect of the changed question order upon the scalogram pattern, it is only necessary to place Board 1 on top of Board 2, and turn the boards over. The change from Board 1 to Board 2, and back again, is accomplished very quickly and in this way changes in scalogram pattern can be observed at all stages of the ordering process. This visual analysis permits one to determine quite easily just how closely the arrangement of respondents from most to least "favorable" conforms to the desired parallelogram picture.

An Example of Scalogram Analysis

Before presenting an actual example in detail, let us briefly summarize the main procedures involved in scalogram analysis.

There are four basic steps in an ordinary scalogram analysis. These are:

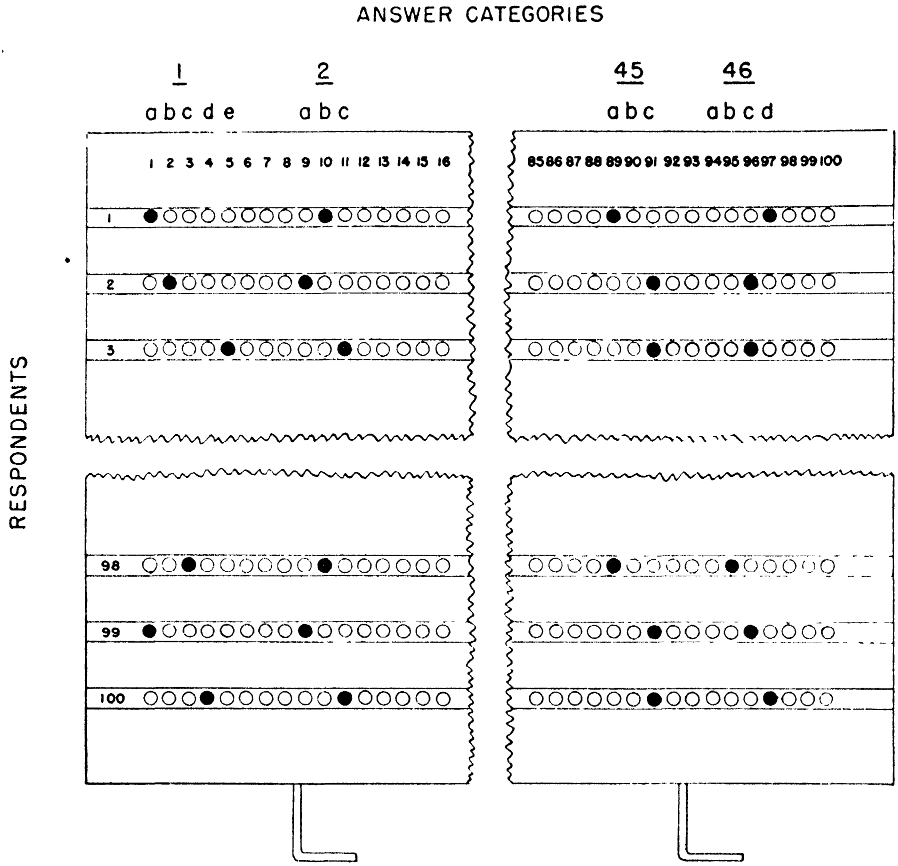


Figure 6. Board 1: Arrangement of individuals. (The filled-in holes represent responses.)

1. *The initial arrangement.* The responses of a sample of 100 individuals to the series of questions to be tested for scalability are tabulated. An initial rank order of questions is determined from the frequencies of "positive" and "negative" responses to each question. The rank order of "positive" response categories runs from the question with the lowest "positive" frequency to the question with the highest "positive" frequency. The rank order

of "negative" response categories runs from the question with the highest "negative" frequency to the question with the lowest "negative" frequency. The rank order of "neutral" responses is unimportant. With the order of response categories established, the answers of the sample of 100 individuals (unordered) are placed into the board, each individual's responses being contained on a single removable slat.

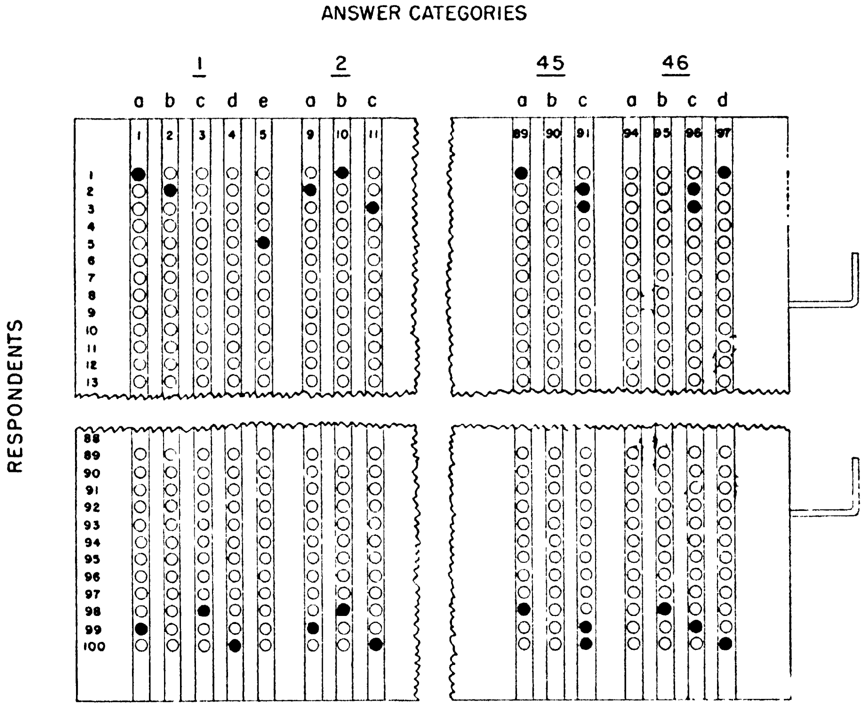


Figure 7. Board 2: Arrangement of questions.

2. *The ranking of respondents.* Respondents are scored for "favorableness" according to the number of "positive" replies they give. The slats are then shifted so that respondents are ranked on the board from "more favorable" to "less favorable." We now have an approximate ranking of respondents from "more" to "less favorable," and an approximate ranking of questions according to frequency of responses to "positive" and "negative" categories. If the data are scalable, the required parallelogram pattern should be apparent at this stage.

3. *Combining categories.* We now transfer the responses from the first scalogram board to the second board. This permits us to shift

answer categories without disturbing the rank order of respondents. The answer categories of each question are shifted so that all categories of a single question or item are adjacent. Those answer categories which overlap each other are combined into a single category. Each item is tested for "improvement" and "error." A new question rank order according to frequency of "positive" and "negative" responses is determined after categories have been combined and, if necessary, questions with many "errors" eliminated. The slats are shifted once more bringing together all "positive" responses and all "negative" responses in proper rank order.

4. *Final arrangement.* Responses are now transferred back to the first board to permit a reshifting of individuals while the rank ordering of questions is held constant. Individuals are once more ranked as in the second step from "more" to "less favorable." The resulting arrangement represents the final scalogram picture. Errors are counted and the coefficient of reproducibility computed.

Now let us take an actual example and carry it through these four stages of scalogram analysis.

Step 1. The Initial Arrangement

A series of questions constituting a sample of an entire universe of items is chosen and given a title indicative of the subject matter covered. The twelve questions used in the present example deal with soldiers' attitudes toward the Army and are given in List 1. The working sample usually consists of 100 cases which are obtained by drawing every "*n*th" schedule or card from the total number of schedules or IBM cards in the survey. By judging its content, the response to each question is designated as "positive," "neutral," or "negative." When several middle positions are present, we may have "positive-neutral," "neutral-neutral," or "negative-neutral" responses. "No answers" are kept distinct from the responses at first. Always, the apparently most favorable responses to a question are designated as "positive," and the least favorable as "negative."

List 1

Attitude toward the Army

(Questions are numbered as they were in the questionnaire.)

26. All things considered, do you think the Army is run about as efficiently as possible, or do you think it could be run better? (Check one)
    1 ______ It is run about as well as possible, everything considered

2 ______ It could be run somewhat better
3 ______ It could be run a lot better

27. Do too many of the things you have to do in the Army seem unnecessary? (Check one)
 1 ______ No, not too many of them seem unnecessary
 2 ______ Yes, too many of them seem unnecessary

28. In general do you think the Army has tried its best to see that men get as square a deal as possible? (Check one)
 1 ______ Yes, it has tried its best
 2 ______ It has tried some but not hard enough
 3 ______ It has hardly tried at all

29. In general do you feel you yourself have gotten a square deal from the Army? (Check one)
 1 ______ Yes, in most ways I have
 2 ______ In some ways, yes, in other ways, no
 3 ______ No, on the whole I haven't gotten a square deal

30. Do you feel that the Army is trying its best to look out for the welfare of enlisted men? (Check one)
 1 ______ Yes, it is trying its best
 2 ______ It is trying some, but not hard enough
 3 ______ It is hardly trying at all

31. In general how interested do you think the Army is in your welfare? (Check one)
 1 ______ Very much
 2 ______ Pretty much
 3 ______ Not so much
 4 ______ Not at all

32. What do you think of the statement that *"The Army makes a man of you?"* (Check one)
 1 ______ There's a lot to it
 2 ______ There may be something to it, but I'm still doubtful
 3 ______ There is not much to it
 4 ______ No opinion

33. In the Army, some jobs are naturally harder and more dangerous than others and the Army has to put men where it thinks they are needed.

 Considering everything, do you think the Army is trying its best to see that, as far as possible, no man gets more than his fair share of the hard and dangerous jobs? (Check one)
 1 ______ Yes, it is trying its best
 2 ______ It is trying some, but not hard enough
 3 ______ It is hardly trying at all

34. Do you think the Army is trying its best to see that the men who have the hard and dangerous jobs get the special consideration and breaks they deserve? (Check one)
 1 ______ Yes, it is trying its best
 2 ______ It is trying some, but not hard enough
 3 ______ It is hardly trying at all

35. On the whole, do you think the Army gives a man a chance to show what he can do? (Check one)
    1 ______ A very good chance
    2 ______ A fairly good chance
    3 ______ Not much of a chance
    4 ______ No chance at all
    5 ______ Undecided

36. In general, how well do you think the Army is run? (Check one)
    1 ______ It is run very well
    2 ______ It is run pretty well
    3 ______ It is not run so well
    4 ______ It is run very poorly
    5 ______ Undecided

37. Do you think when you are discharged you will go back to civilian life with a favorable or unfavorable attitude toward the Army? (Check one)
    1 ______ Very favorable
    2 ______ Fairly favorable
    3 ______ About 50–50
    4 ______ Fairly unfavorable
    5 ______ Very unfavorable

The terms used here like "positive," "negative," and the like, are only to indicate a trial ordering of the categories within each item. They are not meant to imply being on one side or another of a "zero" point. Actually, the scale analysis could be performed "blindly," without looking at the content of the categories. The technique is one of successive approximations which will converge anyhow to the proper scale ordering—if a scale exists. Judging category order within an item beforehand simply enables the first approximation to be a very good one, so that not many more iterations are required. Working "blindly" would involve many more iterations ordinarily. The preliminary ranking of categories by judging their content is not a theoretical necessity; it is only an extremely helpful labor-saving device. If it is uncertain how to rank some categories a priori, which may often be the case with the hyphenated "neutral" categories and with "no answer," these can be given an arbitrary order, which will then be automatically revised in the course of further steps in the analysis.

The pattern of our initial scale setup is determined by the frequencies of the "positive" and "negative" categories of each question. The scalogram is built from left to right, column by column, each column containing a single category of response for a single question. The first column consists of the "positive" category with the fewest responses, and so on until all of the "positive" re-

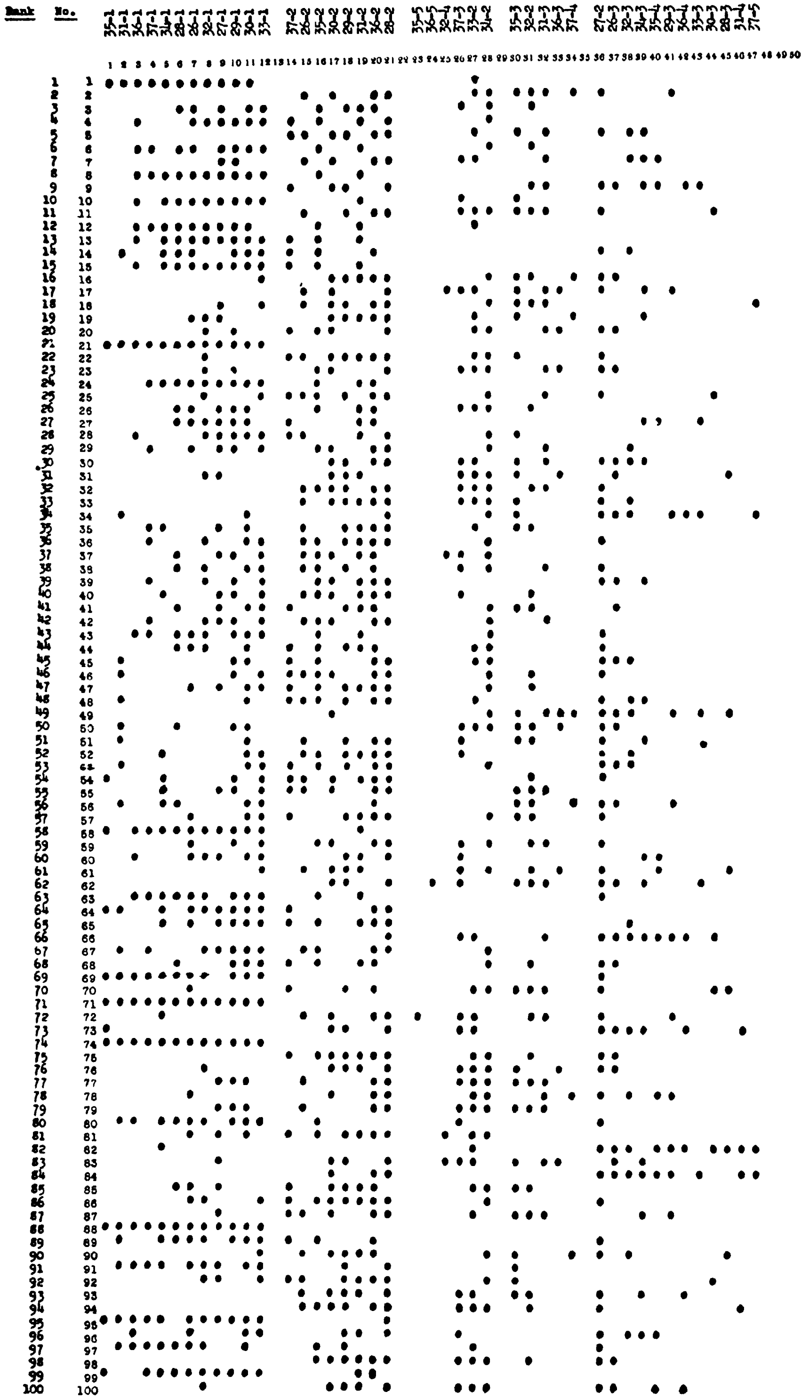
sponses of the questions in the series are represented. Here a column is left vacant to set off the "positive" from the "neutral" responses, which are the next group placed in the board. The "neutral" responses may follow the question, or column, order of either the "positive" or the "negative" responses. If any of the questions have more than one "neutral" response, the "positive-neutral" responses of all the questions are placed first, following the column order assigned the respective "positive" categories. Immediately following are the "neutral-neutrals," and then the "negative-neutrals," following the column order assigned the "negatives." Another vertical column is left vacant and is followed by the "negative" responses. The "negative" category order is determined again on the basis of frequency of "negative" responses. The categories in the "negative" section are placed from left to right in order of highest frequency to lowest. When "no answers" occur, they are placed a few columns to the right of the scale proper. If it is desired to test whether a question of different content correlates with the scale, its responses may be set in further on the right side of the board, with the columns ranging from "positive" to "negative."

Each of the slats holds the answers to every question on the board for one man. Since each horizontal slat bears a number from 1 to 100, and since each schedule or IBM card is numbered, we can readily identify any one of the men for further study or reference.

The scalogram picture for Step 1 shows the initial arrangement. The first column of numbers on the left identifies the rank order of respondents, while the second column identifies the sample of 100 respondents who have been selected at random for scale analysis. The numbers across the top of the scalogram refer to the preliminary order assigned the categories. At the top of the chart the question number and the code number for the response category is given for each column. The frequency for each column is given at the bottom of the chart. Since the total number of answer categories for the twelve questions was 43, only half of the columns of the whole board are pictured. Columns 1 to 12 represent the most "positive" answer category for each of the 12 questions used in this scale. Columns 36 to 47 represent the most "negative" answers for these 12 questions. Between these two columns, we have the intermediate answer categories arranged according to "positive-neutrals" (columns 14 to 21), "neutrals" (columns 23 to 28), and "negative-neutrals" (columns 30 to 34). The frequency of "positive" and

Step 1. Initial Arrangement (Attitude toward Army)

Respondent order *Category order*



"negative" responses for each of the questions in the scale has been tabulated and answer categories have been placed in the board in ascending order of frequency for "positive" categories, and descending order of frequency for "negative" categories. The ordering of "neutral" categories may follow the order of either the "positive" or "negative" categories.

This initial picture, therefore, represents the responses of a random sample of one hundred individuals to twelve questions arranged according to the frequency of "positive" and "negative" answers. The responses of respondent number 57, for example, can be read from the scalogram as follows: slat 57 has balls in columns 7, 11, 12, 14, 18, 19, 20, 21, 28, 30, 31, and 36, which represent the following replies to each of the twelve questions: 26–1, 30–1, 33–1, 37–2, 29–2, 31–2, 36–2, 28–2, 34–2, 35–3, 32–2, 27–2.[7] Similarly, it is possible to see which individuals answered question 35–1: i.e., persons 1, 21, 54, 58, 64, 69, 71, 73, 74, 88, 95, and 99.

An alternative way of arriving at the scalogram for Step 1 would be to use Board 2 in which each slat stands for an answer category and to put one question in the board at a time with all the answer categories for any one question grouped together. The answer categories can be shifted from right to left into the desired question order, and the balls can then be transferred to Board 1 in preparation for Step 2—the ranking of respondents.

Step 2. The Ranking of Respondents

The initial setup completed, we start moving the slats, or respondents, up and down the board to arrange them in rank order of "favorableness." The men who are "favorable" on all twelve questions will have balls in the first twelve columns; these men are moved to the top. After these men come those who are "favorable" on eleven questions, then ten, etc. An approximate method which has proved useful during this initial ranking is to use simple weights, e.g., to give "positive" answers a weight of 4, "positive-neutrals" 3, "neutrals" 2, "negative-neutrals" 1, and "negatives" 0, and then to rank individuals according to their scores. An additional aid is to rank individuals so that as many balls as possible appear in a solid streak for each of the answer categories. As will be shown in Step 3, the number of errors is computed by marking off these "solid streaks" with cutting points on the rank order, and then counting as errors the number of individuals who fall outside these cutting

[7] Code numbers will be used throughout to indicate answer categories.

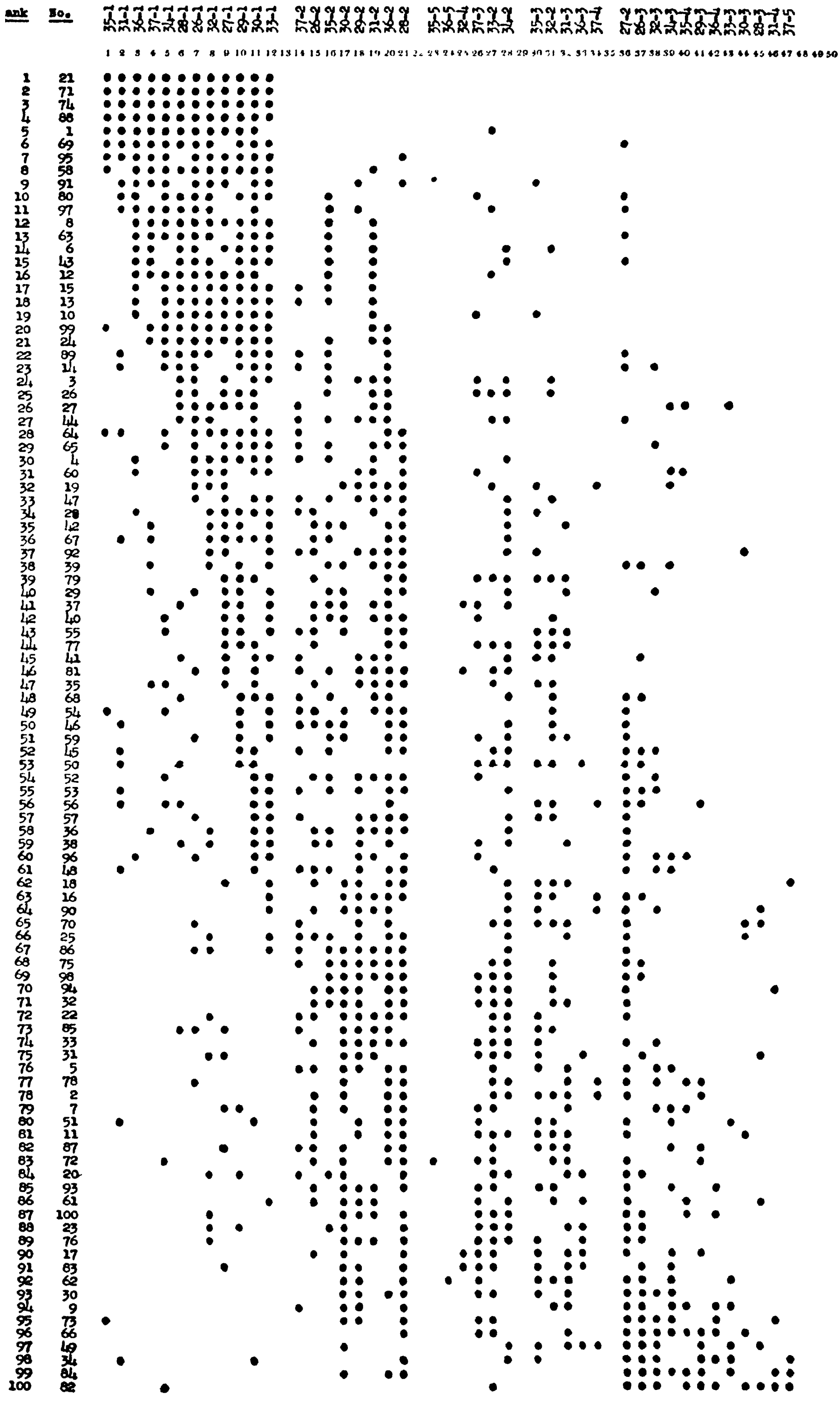
points. The previous chapter has already shown how these "solid streaks" are predictable from a knowledge of the marginal frequency distribution of replies.

Finally, *within each of the scale types separately*,[8] individuals are ranked so that those individuals who are missing balls in the higher numbered columns among the "positives" and in the lower numbered columns among the "negatives" are placed above those individuals who are missing balls not so close to the center columns. The purpose of these arrangements is to secure as many perfect or near perfect scale types as possible. Since the amount of error is computed by counting up deviations from the ideal scale pattern, these criteria of respondent order are aimed at permitting maximum predictability of responses to all twelve questions from a knowledge of the scale score. The basic criterion which determines whether an individual should be shifted up or down is the effect this shift will have upon the number of errors. How errors are computed will be shown in Step 3.

The scalogram for Step 2 shows the results of the first ranking of respondents. Respondents with "more favorable" answers have been moved to the top of the board, while respondents with "less favorable" answers have been moved to the bottom. The first column of numbers on the left shows the rank order of respondents, while the second column of numbers shows the revised arrangement of the respondents' identification numbers which results from the ranking. Answer categories are once more coded at the top of the picture. The rank order at this stage of the arrangement need not be done too scrupulously. The "neutral" replies especially can be handled relatively unsystematically. With practice, the research worker will find that the visual pattern is enough to determine a good initial rank order of individuals. However, let us take several specific examples of why certain individuals were ranked above others to illustrate the general principles involved.

1. "*Weighting.*" Rank orders 9 versus 10 (respondents 91 versus 80). Respondent 91 is placed above respondent 80 because, although both respondents show 9 "positives," person 91 has a weight of 7 in the other 3 categories (two "positive-neutrals" and one "nega-

[8] We define "scale type" for the purpose of scalogram analysis as that perfect scale type which the given individual most closely approaches with the least number of errors. If there is more than one perfect scale type to which the given individual approaches most closely, he is classified as belonging to that scale type which best maintains the "solid streaks."


ank
No.
1 21
2 71
3 74
4 88
5 1
6 69
7 95
8 58
9 91
10 80
11 97
12 8
13 63
14 6
15 43
16 12
17 15
18 13
19 10
20 99
21 24
22 89
23 14
24 3
25 26
26 27
27 44
28 64
29 65
30 4
31 60
32 19
33 47
34 28
35 42
36 67
37 92
38 39
39 79
40 29
41 37
42 40
43 55
44 77
45 41
46 81
47 35
48 68
49 54
50 46
51 59
52 45
53 50
54 52
55 53
56 56
57 57
58 36
59 38
60 96
61 48
62 18
63 16
64 90
65 70
66 25
67 86
68 75
69 98
70 94
71 32
72 22
73 85
74 33
75 31
76 5
77 78
78 2
79 7
80 51
81 11
82 87
83 72
84 20
85 93
86 61
87 100
88 23
89 76
90 17
91 83
92 62
93 30
94 9
95 73
96 66
97 49
98 34
99 84
100 82

tive-neutral") while person 80 has a weight of 5 (one "positive-neutral," one "neutral" and one "negative" response).

2. *"Solid streaks."* Rank orders 52 versus 67 (respondents 45 versus 86). Both respondents 45 and 86 have the same number of "positives," and yet, despite the fact that 45 has a lower total weight (28) than 86 (35), he is placed above person 86. The reason for this arrangement can be seen by looking at column 11 across the top. If person 86 were placed above person 45, this would introduce an empty hole into the solid streak of balls in this column.

3. *"Errors close to center columns."* Rank orders 7 versus 8 (respondents 95 versus 58). Both respondents have the same number of "positives" and the same total weights. However, person 95 is placed above person 58 because the missing ball is in a higher column (column 6) than for 58 (column 2) and, therefore, closer to the center columns.

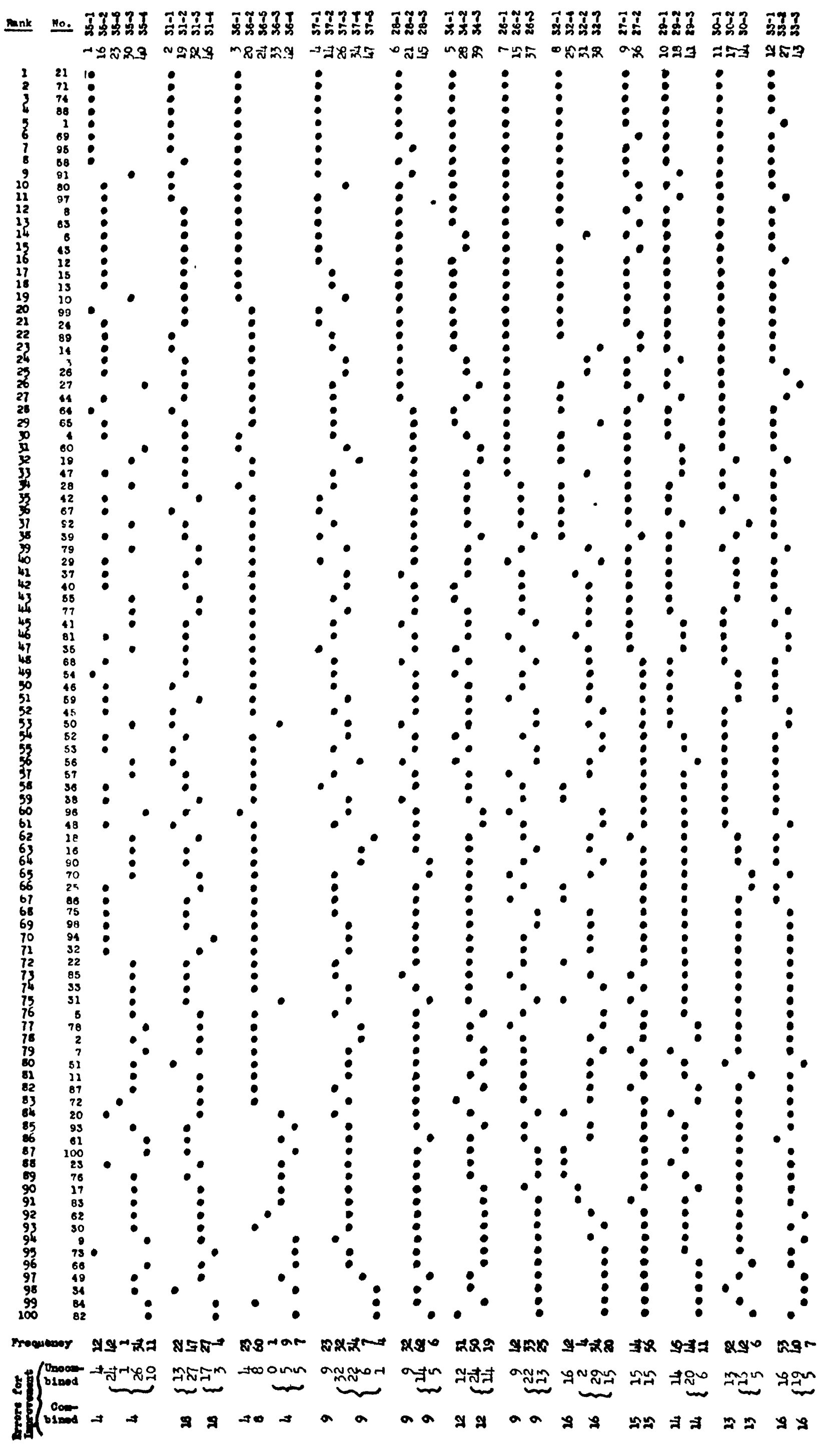
The above examples are intended only to show some of the procedures by which a parallelogram pattern with fewer errors will be obtained. For the most part, this ranking of individuals can be done visually, without scoring. The pattern becomes especially clear when only dichotomous items are used. It is sometimes possible to tell at this very early stage that the items will *not* form a scale, in which case there is no need to proceed with the next steps.

Step 3. Combining Categories

After a first approximate ranking of individuals, the problem becomes one of combining answer categories. As in the case of ranking individuals, the basic criterion for these combinations is the number of errors or deviations from the ideal or expected scale pattern. Since making these combinations involves shifting the answer categories, the balls are transferred to the second board to permit the shifting of answer categories while the rank order of individuals is held constant. Holding the two boards together so that the slats of one are perpendicular to those of the other, we turn them over. The balls fall into Board 2, permitting the shifting of answer categories columns. In order to study each question separately the slats for each individual question are brought together. This is the pattern shown in the scalogram picture for Step 3.

Before proceeding with a specific example, let us review what is meant by the combination of answer categories. Suppose we are ranking 100 individuals in order of "favorableness" based upon

Respondent order *Category order*



their responses to a single question whose marginal frequency distribution is as follows:

Answer category		*Number*
Strongly agree	("Positive")	10
Agree	("Positive-Neutral")	20
Indifferent	("Neutral")	20
Disagree	("Negative-Neutral")	25
Strongly disagree	("Negative")	25
		100

If we assume that this question itself forms a perfect scale, the first ten ranks would be occupied by the respondents who checked the "positive" answer category, the next twenty ranks (from 11 to 30) by the respondents who checked the "positive-neutral" category, and so on. We could mark off these ranks by "cutting points" along the rank continuum. From these cutting points along the rank continuum, we can predict the respondent's answer to the question. For example, a respondent whose rank order is anywhere from 11 to 30 would be predicted as checking the "positive-neutral" category. Now if we hypothesize that this question belongs to a scalable universe, we would expect this prediction of answer category from rank order to be true for each of the single questions in the series of questions which go to make up the scale. Each answer category to each question in the scale will serve to define an additional cutting point. *In a perfect scale,* all responses to the answer category will fall between the cutting points of that category. Any response which falls outside the cutting points will be called an error of *reproducibility*; we would predict the wrong response from knowing that the rank order of the respondent fell between the prescribed cutting points. This is how we compute the error for each of the answer categories shown in the scalogram for Step 3. We mark off the cutting points and then count the number of responses which fall outside the cutting points. Cutting points should mark off consecutive ranks; this means that "solid streaks" in adjacent categories should not overlap and should not omit any ranks. Only *one* cutting point should separate any two adjacent categories.

We now come to an important point. The cutting points should separate answer categories which are distinctly indicative of different degrees of "favorableness" or "unfavorableness." Thus the respondents in the "positive-neutral" category should be distinctly

less "favorable" on the scale continuum than the respondents in the "positive" category, if the rank order is to be reproducible. However, we often find this *not* to be the case. Instead of the rank order of individuals who answer "positive" running from 1 to 10, and the rank order of respondents who answer "positive-neutral" running from 11 to 30, we find the two mixed rather indiscriminately. We could not predict without a large number of errors whether a person whose rank order was from 1 to 30 answered "positive" or "positive-neutral." However, and this is the important point, we *could* predict very well that his answer would be *either* "positive" *or* "positive-neutral." Thus by combining the two categories into one we can set one cutting point between 30 and 31 and predict for the individuals from 1 to 30 that their answers would be either "positive" or 'positive-neutral." This is what is meant by combination—the bringing together of two or more answer categories which cannot be reproduced separately into one combined category which can be reproduced.

The reason for "positive" and "positive-neutral" categories failing to discriminate between meaningful degrees of "favorableness" may be due to the verbal habits of people. Some people who are basically in the same position on the scale continuum will say "strongly agree" where others will say "agree." This problem of verbal habits will be discussed in greater detail in a later analysis of verbal intensity.

Let us look at question 36, for example. The columns 3, 20, 24, 33, and 42 have been brought together so that all the answer categories to question 36 are juxtaposed. The same thing has been done for each of the other questions. We are now ready to make the necessary combinations of answer categories.

Measuring improvement. Whether or not the amount of error indicates that two or more answer categories should be combined can be determined by examining how much the reproducibility of responses would be improved by making the combination.

A way of measuring improvement is to count up the number of errors, both of responses which fall outside the cutting point range and responses which fail to fall within the cutting point range, and to compare this to the number of responses falling within the range (nonerrors). A working rule is that the number of nonerrors should exceed the number of errors for a category to be kept separate.

First, it is necessary to compute the number of errors which occur in each column. Error is determined for each column by counting

up the number of responses in that column which fall outside the cutting points of that answer category, plus the number of responses within these cutting points which do not fall in the column under consideration. Another way of stating this problem would be to ask how well does the rank order of an individual permit one to predict whether this individual checked the answer category for which the error is being computed. For example, we return to the response categories to question 36 in columns 3, 20, 24, 33, and 42. In column 3 the "positive" responses run from ranks 1 to 19 (the first column of figures given on the left indicating the rank order of respondents). This places the cutting point below rank 19, and we now count as errors for improvement any responses which occur below rank 19 or any responses which do *not* occur from 1 to 19 in column 3. This column, therefore, has four errors occurring in ranks 30, 31, 34 and 60. Looking at the "positive-neutrals" to question 36 in column 20, we begin with rank 20 and cont nue to rank 83 to indicate those individuals who in a perfect scale would have replied in this category to question 36. This time we find six errors within this range (in ranks 30, 31, 34, 53, 60, and 75) and two errors falling outside this range (in ranks 93 and 99). Column 33 (category 3) begins at rank 84 and ends at 91, with three responses falling outside this range (ranks 53, 75, and 97) and two errors occurring within it (ranks 85 and 87). Column 24, containing the "neutral" answer to question 36 appearing in rank 92, may be considered as beginning and ending at rank 92, thus showing no errors within this range. It is not necessary that the "neutral" answer be placed between the "positive-neutrals" and the "negative-neutrals," if other placement will reduce the overall scale error. Column 42 begins at rank 93 and runs to 100, showing three errors within (ranks 93, 97, and 99) and two errors outside (ranks 85 and 87) this range. In this manner the errors are computed for all columns. These errors are given at the bottom of Scalogram 3 in the row marked "Errors for Improvement—Uncombined."

We are now ready to see if certain answer categories should be combined. To tell this we examine each column and decide whether the category is sufficiently reproducible to stand by itself, or whether the pattern of errors indicates that it does not permit reproducibility unless combined with the next more or less "favorable" category. The criterion to be applied is: to what extent will the combination of answer categories increase reproducibility from the rank order of the respondents? The problem is essentially one of finding out

whether to combine the "neutral" categories, "positively" or "negatively," if at all. As was discussed in the previous chapter, there are certain advantages to keeping as many categories intact as possible. Additional categories increase the number of scale types between which the sample of items can discriminate. Furthermore the possibility that a sample of questions chosen from a nonscalable universe will apparently form a scale is greatly reduced as the number of answer categories that can be kept separate increases.

A practical procedure is to count the number of errors in each category, and if this number is relatively small (e.g., if it is less than the number of nonerrors) then the category may be kept separate. If the category has too many errors to stand by itself, then an attempt is made to reduce this error by combination with an adjacent category. If no combinations will reduce the number of errors below that of the nonerrors, then the conclusion is that the question is not a simple function of the scale score. This indicates the presence of one or more additional variables and does not permit a rank ordering of individuals on a single continuum.

Once again, let us look at question 36 for a specific example. Category 1 (column 3) can stand by itself since it makes only 4 errors of reproducibility from the rank order as compared with 19 nonerrors. Similarly, category 2 (column 20) makes only 8 errors compared to 58 nonerrors and can also stand by itself. Category 5 (column 24) contains only one case and so should be combined. Combining it with category 2 serves only to increase the amount of error in 2, whereas it can be combined without increasing errors with category 3. Category 3 (column 33) has 5 errors, but contains only 6 nonerrors, and so it just barely satisfies our criterion. Looking at category 4 (column 42) we see that this category could not stand by itself (five errors and five nonerrors). However, combining categories 3 and 4 and treating responses to these two categories as if they were a single category produces only 5 errors compared to 14 nonerrors. Combining category 5 with categories 3 and 4 raises the number of nonerrors to 15 and reduces the number of errors to 4. The combination of categories 3, 4, and 5, therefore, reduces the number of errors from 10 to 4.

By the same process of reasoning, we make combinations for all the other questions in the scale as shown in the last row of the scalogram picture for Step 3. By visual analysis of the overlapping of balls within the answer categories of each question, it can be determined very rapidly how best to combine the categories so as to

minimize the ratio of errors to nonerrors. To combine columns it is necessary only to remove the balls from the columns to be combined with a magnet and to transfer them all to a single column.

An especially good example of how scalogram analysis aids in the determination of how to combine answer categories, and incidentally indicates the weakness of assuming that each answer category actually represents a meaningful change in degree, can be seen from question 37. The balls in the answer categories 2, 3, 4, and 5, despite the apparently different manifest intensity of the categories, are so overlapping as to belong together. This would seem to indicate that the respondents are not interpreting the different answer categories as representing degrees of "favorableness," as evidenced by the fact that a person's answers cannot be reproduced adequately from his rank order on the whole scale. In the same way it is possible, for example, to see that the "neutral" category in question 34 should be combined with the "negative" category. When the decision on how to combine answer categories is an important one, it is often desirable to increase the sample of respondents to about two hundred, to provide greater stability from the point of view of sampling of people.

Another principle to be observed in making combinations is that of the differentiation of new scale types. As discussed in the previous chapter, the number of scale types depends upon the number of different cutting points on the rank order. Two questions with the same marginal frequencies of responses will cut the scale rank order in approximately the same place and therefore not contribute to the ability of the scale to discriminate. It may sometimes be worth while to increase the error slightly in order to secure a combination of categories which would produce different marginal frequency distributions. The goal is to secure a series of questions which permit the widest range of expression from 0 to 100 per cent "favorable."

Combining answer categories results in a change in "positive" and "negative" frequencies and therefore the categories or columns must be reordered. A more rigorous use of the scalogram board would require that question order and respondent order be checked after each single combination made. However, it has been found that completing all combinations at once and then reordering produces satisfactory results. When the category or column order has been reordered with the "positive" categories ranging from low to

high frequency, and the "negative" categories from high to low (as in Step 1), the balls are ready to be returned to the first board for the final step.

Step 4. Final Arrangement

The combination and reordering of categories now requires a change in the rank ordering of individuals. To shift individuals up and down, it is necessary to return the balls to the first board. The shifting of respondents is done in exactly the same way as in Step 2. Respondents with "more favorable" answers are moved to the top of the board, while respondents with "less favorable" answers are moved to the bottom. Once again the basic criterion which determines whether an individual should be shifted up or down is the number of errors which result from the shift. Due to the decreased number of answer categories, the effect of any shift upon the number of errors is much more easily discernible than in Step 2. The same criteria of "weighting," "solid streaks," and "closeness of error to center columns" serve as aids to the ranking of individuals.

One final check serves to help obtain the least amount of error. The ideal scale types are marked off by means of the cutting points for each column and each person is checked to see if moving him to a lower or higher scale position would reduce the number of errors by which he deviates from the ideal scale type. In general, if a respondent can fit equally well into any of several scale types, he is put into that scale type closest to the middle of the rank order.

The method of determining ideal scale types has been discussed in the previous chapter. The number of types consists of the number of answer categories, minus the number of questions, plus one. In the present example this would be:

$$(2 + 2 + 3 + 2 + 2 + 2 + 2 + 2 + 2 + 2 + 2 + 2) - 12 + 1 = 14$$

Perfect scale types would be represented as in Figure 8.

The scalogram for Step 4 shows the obtained pattern after the individuals have been reordered following the combination and reordering of categories in Step 3. Comparison of the obtained parallelogram with the ideal parallelogram permits one to see clearly the errors of reproducibility. Note that the reordering of individuals changes the number of errors for each category from that of the scalogram for Step 3.

Scale	Question	35	37	36	34	28	26	32	27	29	30	33	31	36	35	37	34	28	26	32	27	29	30	33	31	36
type	Category	1	1	1	1	1	1	1	1	1	1	1	1,2	2	2–5	2–5	2,3	2,3	2,3	2–4	2	2,3	2,3	2,3	3,4	3–5
13		X	X	X	X	X	X	X	X	X	X	X	X													
12			X	X	X	X	X	X	X	X	X	X	X		X											
11				X	X	X	X	X	X	X	X	X	X		X	X										
10					X	X	X	X	X	X	X	X	X	X	X	X										
9						X	X	X	X	X	X	X	X	X	X	X	X									
8							X	X	X	X	X	X	X	X	X	X	X	X								
7								X	X	X	X	X	X	X	X	X	X	X	X							
6									X	X	X	X	X	X	X	X	X	X	X	X						
5										X	X	X	X	X	X	X	X	X	X	X	X					
4											X	X	X	X	X	X	X	X	X	X	X	X				
3												X	X	X	X	X	X	X	X	X	X	X	X			
2													X	X	X	X	X	X	X	X	X	X	X	X		
1														X	X	X	X	X	X	X	X	X	X	X	X	
0															X	X	X	X	X	X	X	X	X	X	X	X

Figure 8. Perfect scale types.

Test for Scalability

We are now ready to test the scalogram for the existence of a satisfactory scale. Four main criteria are used: (1) the coefficient of reproducibility, (2) the number of items and response categories, (3) the range of marginal frequencies, and (4) the pattern of error.

1. *The coefficient of reproducibility.* The coefficient of reproducibility is computed by means of the following formula:

$$\begin{aligned} \begin{array}{l}\text{Coefficient of}\\ \text{Reproducibility}\end{array} &= 1 - \frac{\text{number of errors}}{\text{number of questions} \times \text{number of respondents}} \\ &= 1 - \frac{134}{12 \times 100} \\ &= .89 \end{aligned}$$

For computing the coefficient of reproducibility, only one response is counted for each question. A single response to a question is reproduced from the scale score either correctly or incorrectly. This means that only responses which occur outside the scale type are counted as errors. It will be remembered that in computing errors for improvement (Step 3) the absence of responses *within* the scale type and the occurrence of responses *outside* the scale type were both counted. For reproducibility, errors are counted only once, and the easiest way to spot these errors is to count the number of balls outside the cutting points. The number of errors for each column are shown in the last row of the scalogram for Step 4.

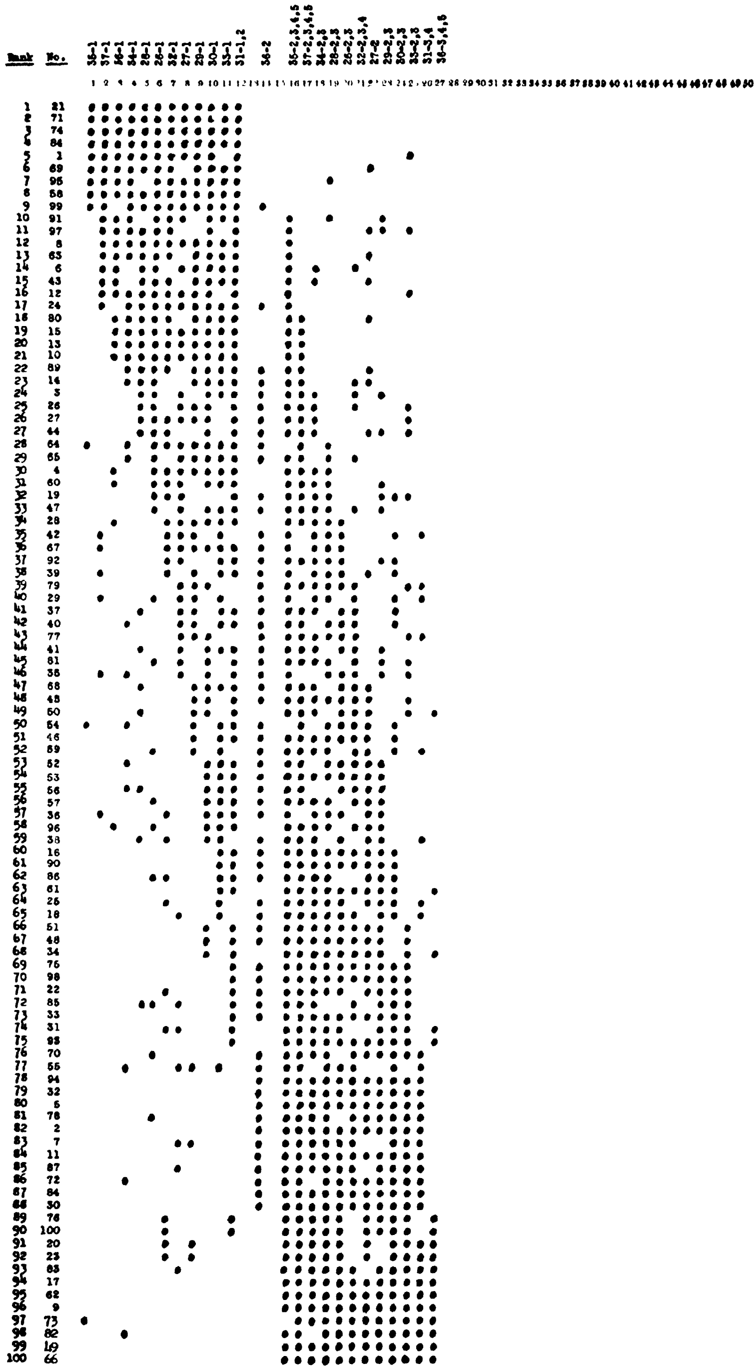
As discussed in the previous chapter, the lower acceptable limit for the coefficient of reproducibility has been placed at about .90.

The coefficient of reproducibility, while a *necessary* test of the existence of a scale, is not a *sufficient* one if only a sample of items is used. These other conditions must be considered: the number of answer categories left after combination, the item marginal frequencies, and the pattern of error.

2. *The number of answer categories.* If more than ten items are used in the sample, as in the present example, then ordinarily the final number of categories does not matter; all items might be dichotomized. However, the more categories that can remain uncombined, the more credible is the inference that the universe is scalable. If less than ten items are used, it may not be too safe to infer that the universe is scalable if all the items must be dichotomized in order to obtain high reproducibility; at least some should be

STEP 4. FINAL ARRANGEMENT (Attitude toward the Army) (Coefficient of Reproducibility = .89)

Respondent order *Category order*



retainable in trichotomous form to make the inference plausible. It may well happen that a sample of items from a nonscalable universe might yield high reproducibility when dichotomized, if the sample is not very large. Of course, once scalability is established, then any further combinations can be made to reduce the number of scale types to be used in practice, if that is desired. In the actual test for scalability, however, it is best to keep as many answer categories as possible.

3. *Range of marginal frequencies.* With respect to the criterion of item marginals, it is essential to remember that reproducibility may be artificially high because the items have extreme kinds of frequencies. To avoid reproducibility being artificially high, some dichotomies should have marginals close to 50–50. Similarly, for trichotomies and other multicategoried questions, the great bulk of the respondents should not be in any one category (after combination), except possibly for an occasional question. The present example with its wide range of marginal frequencies, half of which lie between 40 and 60, satisfies this criterion very well.

4. *Pattern of error.* The pattern of error should be inspected to see that there are no substantial nonscale types of persons. Solid segments in a column that fall outside the cutting points indicate the presence of definite additional major factors in the responses, so that more than one appreciable dimension is present. Nonscale types can be recognized from the occurrence of the same kind of error for a large (i.e., five or more) number of respondents. None are obvious in the present case.

It is also important to remember that scale analysis should not be depended upon to determine content. An item of differing content may fit into the scale pattern of an area, while items with homogeneous content need not scale. Scale analysis can provide some auxiliary aid in interpreting content especially in pretests, but it is important to keep in mind the cautions discussed previously.

From the scalogram picture for Step 4 it is now possible to determine how the entire population should be scored. As discussed previously, a simple weighting scheme is sufficient. Categories of dichotomized items are weighted 1 and 0; categories of trichotomized items are weighted 2, 1, and 0; etc. A more rigorous method would involve scoring individuals according to the ideal scale type which fits with the least amount of error, rather than by weighting categories. For example, individual 69 (rank order 6) would be scored 13 according to closest ideal scale type, but 12 according to weights.

To score this individual in ideal scale type 13 (positive answers to all twelve questions) would result in only one error (his negative response 27–2). However, scoring him as 12 (the number of positive responses made) results in his deviation from the ideal scale type 12 on two counts—his positive response 35–1 and his negative response 27–2 (see Figure 8). A more extreme example would be individual 20 (rank order 91) who fits into ideal scale type 0 according to the least error criterion (two errors of reproducibility, 32–1 and 29–1), but who would receive a weighted score of 2 (increasing errors of reproducibility from ideal scale type 2 to four errors—32–1, 29–1, plus 31–3, 4, and 36–3, 4, 5). While scoring according to closest ideal scale type is desirable, it is a very slow process. The correlation between the two methods is so high as to indicate the acceptability of the weighting method for practical purposes.

A simple scoring diagram for a dichotomous scale when IBM card-sorting equipment is available would take the following form, as illustrated in Figure 9 for the first five dichotomies from the present example.

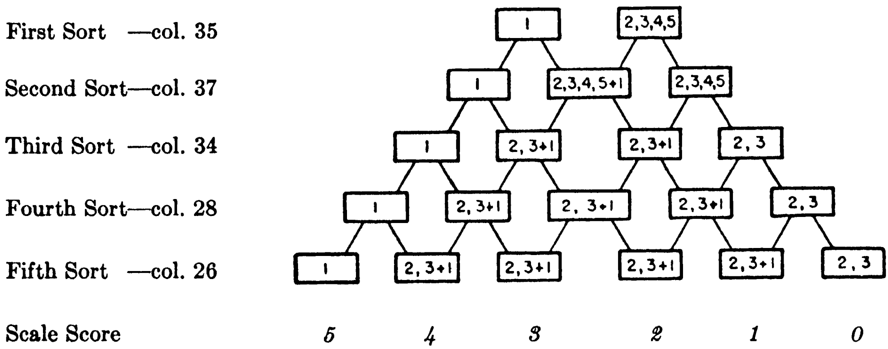


Figure 9. Scoring diagram for dichotomies.

Scaling with only one board. With experience in recognizing the parallelogram pattern and errors of reproducibility, the entire scalogram procedure can be carried out on one board only. This has an important economic advantage, since the cost of constructing two perfectly matched boards can be eliminated. The construction of one board alone composed simply of one hundred removable slats with one hundred holes[9] in each slat can be done very

[9] Less than 100 or more than 100 holes can be used, since the single board need no longer be square. Also since no transfer of balls from board to board is necessary, the holes need not be drilled so exactly.

cheaply. With a grasp of the principles and methods involved in the use of two boards, the one-board method becomes quite feasible.

This method consists of Steps 1 and 2 exactly as outlined. However, Step 3, requiring the combination of answer categories and the reordering of questions, is also carried out on Board 1. This is done by inspection of each question separately as in Step 3, except that the different answer categories of each question remain in their original position instead of being juxtaposed. Thus the combinations for question 36, for example, are decided upon by looking at columns 3, 20, 24, 33, and 42 in Step 2 and computing the error as outlined in Step 3. The combination decided upon is written down for each question, the new frequencies of "positive" and "negative" are computed, the board is cleared, and Steps 1 and 2 are repeated, this time using the new question frequency order. Step 4 is then completed exactly as outlined. Step 3 is the only step involving the use of a second board. In the case of dichotomous items where there is no problem of combining answer categories, one board is equally as satisfactory as two. This method using only one board, while not as manipulatable as the two-board method, has been used very satisfactorily in practice.

The amount of work. The scalogram for Step 4 shows the final multivariate pattern of responses. Some of the various functions it can serve will be discussed in the following chapter. Other uses will probably be forthcoming with additional research. It cannot be emphasized too strongly that this technique is still young, so that all of its potential weaknesses and strengths are as yet far from known. Undoubtedly many of the specific procedures outlined above can and will be improved upon. The best claim that can be made for the present procedure is that it has worked and proved its utility over four years of attitude and opinion research in the Army. In practice it has been found that a clerk can be trained fairly easily to carry through from Step 1 to Step 4 in about eight hours.

CHAPTER 5

THE UTILITY OF SCALOGRAM ANALYSIS[1]

It is important to remember that, in many cases, the present research represents only first approximations to perfect scales. Once an area was found to have sufficient reproducibility, little attempt was made to study it more carefully by means of additional pretesting. In fact, in many cases it was often necessary, due to the press of wartime conditions, to treat an area as if it were scalable even though the four criteria of scalability were met only imperfectly. It was also found necessary in many instances to omit questions from an area because of low coefficients of improvement, without making an additional pretest to find out what extra factors were involved in the questions. The following examples are therefore included as illustrations of a technique rather than for their specific content value.

The scale pictures, or scalograms, on the following pages present briefly several actual examples of scale analysis utilizing the scalogram board technique. Each scalogram will list the questions used and the combinations made in the answer categories. The extent to which the four criteria of scalability are met will be indicated in each case. Following each scalogram will be a brief comment on some of the uses which were made of this scalogram. A detailed discussion of the utility of scale analysis in general will be given in the second section of this chapter.

[1] By Edward A. Suchman.

SECTION I

EXAMPLES OF SCALOGRAM ANALYSIS

The following scalograms are presented for discussion:

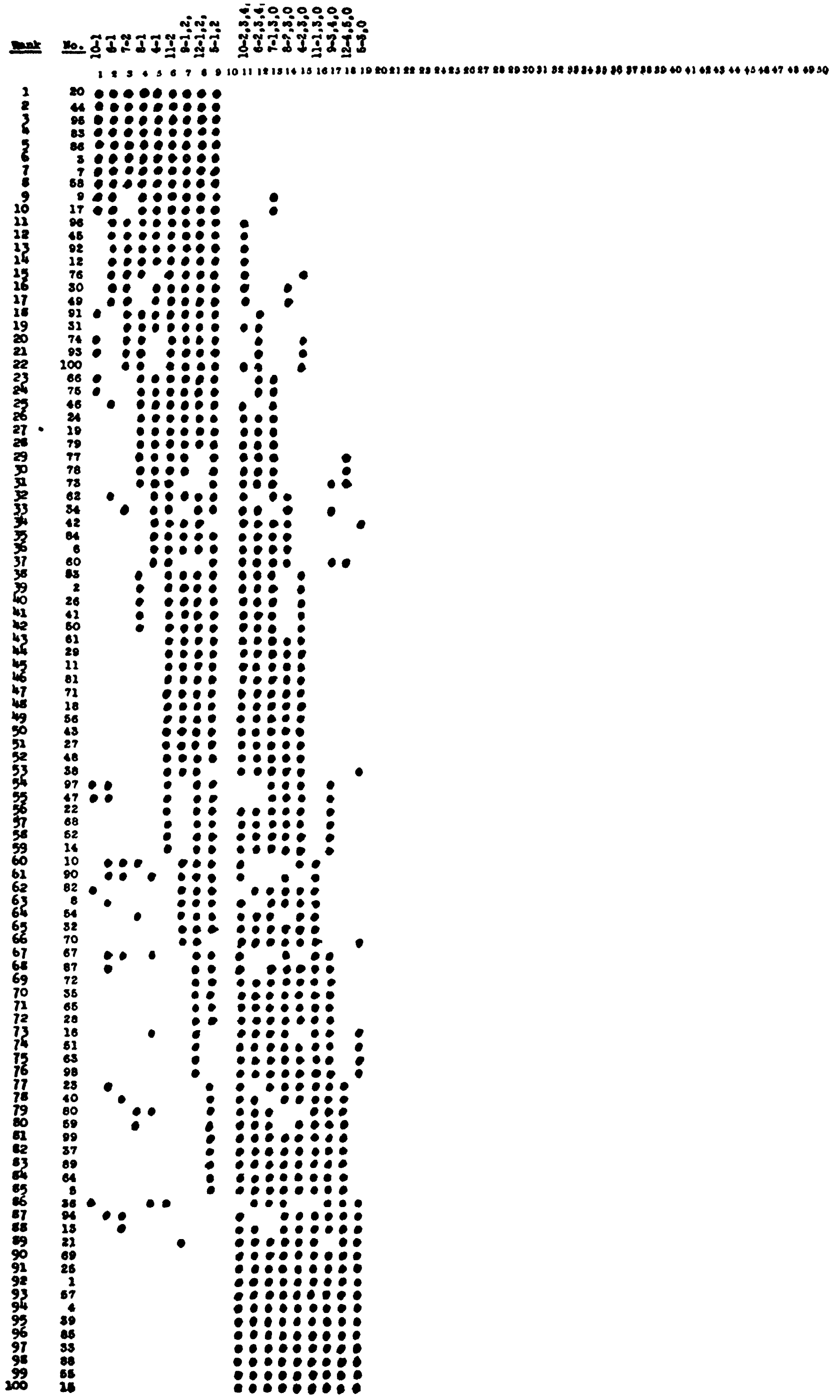
Scalogram	*Universe*
1	Satisfaction with One's Army Job
2	Attitude toward Officers
3	Attitude toward the WAC
4	Attitude toward Postwar Conscription
5	Attitude toward an Army Career
6	Knowledge of Current Events
7	Fear Symptoms

1. *Job satisfaction.* The content of this scale, as can be seen from the specific questions asked, concerns the extent to which the soldier is satisfied with his job in the Army. The sample consists of nine dichotomized questions and has a coefficient of reproducibility of .92. Errors appear to be largely random, with no sizable nonscale types. The range of positive frequencies for the different questions is wide, running from 19 per cent to 78 per cent. One question (question 11, columns 6 and 16), "Do you usually feel that what you are doing in the Army is worthwhile or not?" shows almost perfect reproducibility (only 1 error in rank 86). Respondents in ranks from 1 to 59 reply, "I usually feel it is worthwhile," while respondents in ranks 60 to 100 reply, "I usually feel it is not worthwhile" or "Undecided." This single question, therefore, can serve as a good approximation to the entire scale if a simple division of the population into two ranks is desired with about 60 per cent in the upper rank and 40 per cent in the lower rank. If it is desired to split the population into some ratio different from 60 : 40, the question which has the marginal frequency with the desired ratio is the one to use, even though its reproducibility may not be perfect, if a single question is to be used at all. In the case of imperfect reproducibility it is, of course, desirable to use more than one question, even if only a dichotomization of the population is wanted.

2. *Attitude toward officers.* This scalogram deals with the enlisted man's opinion of his officers. The sample of items consists of eleven questions, all dichotomized. The scalogram has a coefficient of reproducibility of .90. The range of marginal frequencies runs from 8 per cent favorable to 79 per cent favorable, and errors are random. In addition to these eleven questions, the re-

SCALOGRAM No. 1. SATISFACTION WITH ONE'S ARMY JOB (Coefficient of Reproducibility = .92)

Respondent order *Category order*



Scalogram No. 1. Satisfaction with One's Army Job

Questions and Answer Categories

4. How do you feel about the importance of the work you are doing now as compared with other jobs you might be doing in the Army?
   1 ______ It is as important as any other job I could do
   2 ______ It is fairly important, but I could do more important work
   3 ______ It hardly seems important at all

5. How interested are you in your Army job?
   1 ______ Very much interested
   2 ______ A little, but not much
   3 ______ Not interested at all

6. How satisfied are you about being in your present Army job instead of some other Army job?
   1 ______ Very satisfied
   2 ______ Satisfied
   3 ______ It does not make any difference to me
   4 ______ Dissatisfied
   5 ______ Very dissatisfied

7. Would you change to some other Army job if given a chance?
   1 ______ Yes
   2 ______ No
   3 ______ Undecided

8. Do you feel that everything possible has been done to place you in the Army job where you best fit?
   1 ______ Yes
   2 ______ No
   3 ______ Undecided

9. Do you consider your own present job or duty in the Army an important one in the war effort?
   1 ______ Very important
   2 ______ Pretty important
   3 ______ Not so important
   4 ______ Not important at all
   5 ______ Undecided

10. On the whole, do you think the Army is giving you a chance to show what you can do?
   1 ______ A very good chance
   2 ______ A fairly good chance
   3 ______ Not much of a chance
   4 ______ No chance at all
   5 ______ Undecided

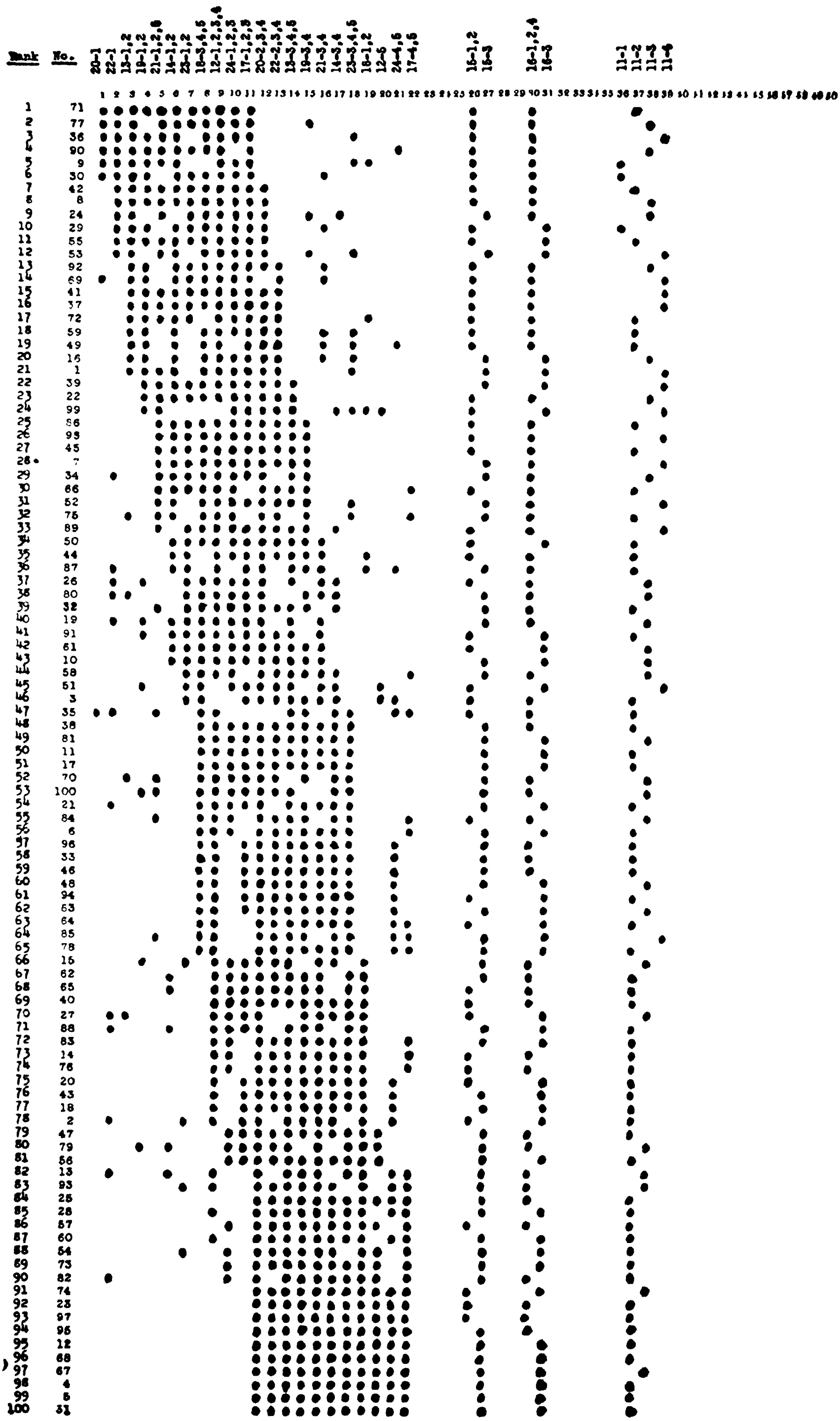
11. Do you usually feel that what you are doing in the Army is worthwhile or not?
   1 ______ I usually feel that *it is not* worthwhile
   2 ______ I usually feel *it is* worthwhile
   3 ______ Undecided

12. Which of the following would you say best applies to your job?
   1 ______ Time always passes quickly
   2 ______ Time passes quickly most of the time
   3 ______ Enjoy working part of the time, but it drags at other times
   4 ______ Time drags most of the time
   5 ______ Time always drags

"No answers" are all coded 0.

Scalogram No. 2. Attitude toward Officers (Coefficient of Reproducibility .90)

Respondent order *Category order*



Scalogram No. 2. Attitude toward Officers

Questions and Answer Categories

11. How do you feel about the privileges that officers get compared with those which enlisted men get?
    1 ______ Officers have *far too many* privileges
    2 ______ Officers have *a few too many* privileges
    3 ______ Officers have *about the right number* of privileges
    4 ______ Officers have *too few* privileges

12. Did your officers give you a good chance to ask questions as to the reason why things were done the way they were?
    1 ______ Yes, always
    2 ______ Yes, usually
    3 ______ Undecided
    4 ______ No, not very often
    5 ______ No, almost never

13. How many of your officers took a personal interest in their men?
    1 ______ All of them
    2 ______ Most of them
    3 ______ About half of them
    4 ______ Few of them
    5 ______ None of them

14. Do you think that your officers generally did what they could to help you?
    1 ______ Yes, all the time
    2 ______ Yes, most of the time
    3 ______ No, they often did not
    4 ______ No, they almost never did

15. How well do you feel that your officers understood your problems and needs?
    1 ______ They were very much aware of my problems and needs
    2 ______ They were fairly well aware of my problems and needs
    3 ______ They did not know very much about my real problems and needs

16. Do you feel that your officers recognized your abilities and what you were able to do?
    1 ______ Yes, I'm sure they did
    2 ______ Yes, I think they did, but I'm not sure
    3 ______ No, I don't think they did
    4 ______ Undecided

17. In general, how good would you say your officers were?
    1 ______ Very good
    2 ______ Fairly good
    3 ______ About average
    4 ______ Pretty poor
    5 ______ Very poor

18. How many of your officers used their rank in ways that seemed unnecessary to you?
    1 ______ Almost all of them
    2 ______ Most of them
    3 ______ Some of them
    4 ______ Only a few of them
    5 ______ None of them

19. When you did a particularly good job did you usually get recognition or praise for it from your officers?
    1 ______ Always
    2 ______ Usually
    3 ______ Rarely
    4 ______ Never

20. How much did you personally like your officers?
    1 ______ Very much
    2 ______ Pretty much
    3 ______ Not so much
    4 ______ Not at all

21. How did you feel about the officers that had been selected by the Army?
    1 ______ There were the best ones that could have been selected
    2 ______ They were as good as any that could have been picked
    3 ______ Somewhat better ones could have been picked
    4 ______ Much better ones could have been picked
    5 ______ Undecided

22. How much did you personally respect your officers?
    1 ______ Very much
    2 ______ Pretty much
    3 ______ Not so much
    4 ______ Not at all

23. On the basis of your Army experience, do you think relations between officers and enlisted men were satisfactory or unsatisfactory?
    1 ______ Very satisfactory
    2 ______ Fairly satisfactory
    3 ______ Undecided
    4 ______ Fairly unsatisfactory
    5 ______ Very unsatisfactory

24. When you are discharged from the Army, do you think you will go back to civilian life with a favorable or unfavorable attitude toward the officers in the Army?
    1 ______ Very favorable
    2 ______ Fairly favorable
    3 ______ About 50-50
    4 ______ Fairly unfavorable
    5 ______ Very unfavorable

sponses to three other questions are given on the side of the scalogram. Two questions, numbers 15 and 16, are offered as examples of items which did not conform to the scale pattern. They were originally included as items characterizing a soldier's attitude toward his officers, but, as can be seen from the scalogram picture, contain a large number of errors. One cannot predict well enough from a knowledge of the respondent's scale score, based upon his answers to the other eleven reproducible items, what his responses to these two questions would be.

The next step in a complete scale analysis would be to pretest these two questions further in an attempt to find out what additional variables they contain. By means of such a pretest it may

be possible to break down the universe of attitude toward officers into further subuniverses. Or it may be decided that the additional variables are irrelevant and one can proceed with the eleven items that were scalable. It should not be surprising that acceptable scales may not be forthcoming from the first analysis of a series of items. Most social attitudes are complex and it is to be expected that they contain more than one variable.

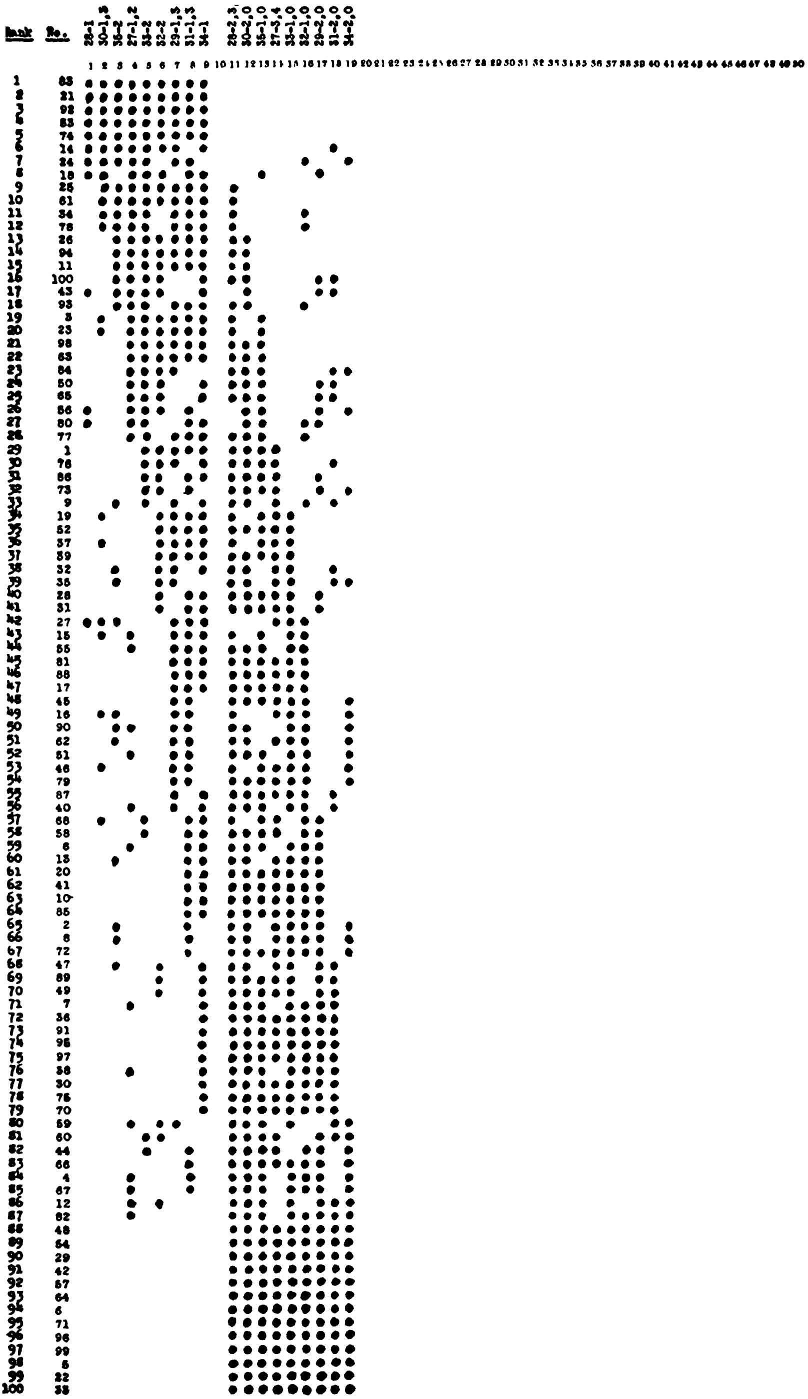
The third question on the side, number 11, was not intended as part of the universe of attitude toward officers, but was included in the scalogram picture merely to see how closely it correlated with the attitude items. If it did happen to conform to the scale pattern, this might mean that the universe to which it belonged had a close relationship with attitude toward officers. However, the present scalogram alone is not sufficient to establish the size of the correlation; additional items similar in content to the one in question would have to be constructed and also tested along with the present series. It should especially be remembered that even if an item is reproducible perfectly from scale scores, this is not proof that the item is part of the definition of the universe. No matter how item 11 correlated with attitude toward officers, we would not on the basis of this correlation consider it as part of that universe. Only a judgment of content can determine what belongs in a universe, and not correlations or reproducibility. This problem of determining the items of a universe will be discussed in greater detail in the next section.

3. *Attitude toward the WAC.* The scalogram picture of attitudes of enlisted men toward the Women's Army Corps is shown in Scalogram 3. The sample is composed of nine dichotomized questions dealing with various aspects of the WAC and has a coefficient of reproducibility of .89. Marginal frequencies are distributed evenly over a range from 12 per cent "favorable" to 64 per cent "favorable." Question 35 (columns 3 and 13) is of low reproducibility. Column 3 shows 12 errors and 17 nonerrors. In addition, there is a sizable number of errors in the "positive" answer category (column 3) which cluster together in the low ranks, indicating that there is a group of men with relatively "unfavorable" attitudes toward the WAC who nevertheless answer "favorably" on this question. This shows that the errors, in addition to being large in number, are not random. What possible interpretation is there of this nonscale question?

One hypothesis, *which requires further testing before it can be ac-*

SCALOGRAM NO. 3. ATTITUDE TOWARD THE WAC (Coefficient of Reproducibility = .89)

Respondent order *Category order*



cepted, is that the second variable which this question introduces into the scale is one of factual information about the WAC, and not of attitude or opinion. The question asks, "Do you agree or disagree with the following statement, 'WAC's get ratings a lot easier than men do?' " In general men who are relatively favorable to

Scalogram No. 3. Attitude toward the WAC

Questions and Answer Categories

27. In your opinion, how necessary is it for the war effort to have WAC's in the Army?
    1 ______ Very necessary
    2 ______ Pretty necessary
    3 ______ Not so necessary
    4 ______ Not necessary at all
    5 ______ Undecided

28. Suppose a girl friend of yours was considering joining the WAC. Would you advise her to join or not to join?
    1 ______ I would advise her to join
    2 ______ I would advise her not to join
    3 ______ Undecided

29. In your opinion are the jobs which women in the WAC do more important or less important than the jobs which are done by men in the Army who are not on combat duty? Or are they equally important?
    1 ______ The WAC does more important jobs
    2 ______ The WAC does less important jobs
    3 ______ The jobs are equally important

30. A woman does more for her country in the WAC than she can do by working in a war industry.
    1 ______ Agree
    2 ______ Disagree
    3 ______ Undecided

31. The training a woman gets in the WAC will be useful in civilian life.
    1 ______ Agree
    2 ______ Disagree
    3 ______ Undecided

32. Being a WAC is bad for a girl's reputation.
    1 ______ Agree
    2 ______ Disagree

33. The WAC is no place for a girl.
    1 ______ Agree
    2 ______ Disagree

34. WAC officers deserve a salute just the same as men officers.
    1 ______ Agree
    2 ______ Disagree

35. WAC's get ratings a lot easier than men do.
    1 ______ Agree
    2 ______ Disagree

"No answers" are all coded 0.

the WAC on other questions are inclined to disagree with this statement, showing that this item, to a certain extent, does involve the variable "attitude toward the WAC." However, many men who in general are among the least favorable to the WAC are also inclined to disagree with this statement despite their generally "unfavorable" attitude. It may be that these men are answering the question based on their knowledge that, as a matter of fact, WAC's do *not* get ratings a lot easier than men do. This hypothesis can only be checked by additional pretesting.

Another interesting example of the use of scalogram analysis in pretesting the meaning of questions is also given by this scale. In a pretest form, question 27 originally read, "In your opinion how necessary is it for the war effort to have *women* in the Army?" as compared to the present wording which substitutes the word "WAC's" for "women." It was found, upon testing responses for scalability, that this question as originally worded contained a great deal of error. Many individuals who occupied low ranks on the scale answered that it *was* necessary to have "women" in the Army. Further pretesting revealed the reason for this discrepancy. These individuals were interpreting the "women" to include nurses rather than only WAC's. Substituting the word "WAC's" for "women," removed this source of error as can be seen from column 4 of the present scalogram picture.

4. *Attitude toward postwar conscription.* This is only a five-question scale, but it contains two trichotomies, which help to reduce the possibility that this sample of five questions could produce the scale pattern and yet not belong to a scalable universe. Reproducibility is .94, errors are random, and the range of marginals is diverse.

The two trichotomies in this scale illustrate quite clearly what is implied by a separate middle response category. For example, the first trichotomy, based on question 60, divides the population into three ranks with only seven errors of reproducibility. In 93 cases out of 100, we could predict correctly how the respondent answered this question on the basis of the respondent's rank order. This single question could be used alone to divide the population into three rank groups with regard to the issue of universal military training. The fact that this question was selected from the scalogram pattern gives one increased assurance that, even if a different question were asked from the indefinite number of opinion questions that could be asked on this issue, the rank ordering of individuals would be essentially the same as on this question.

It is also instructive to note what these two trichotomies contribute to the scale continuum in the way of differentiating scale types. These scale types can be marked off according to the procedure previously outlined. The number of scale types would be decreased from eight to six if these trichotomies were to be dichotomized; combining the middle category of question 60 with the top category (column 6 with column 2) would combine scale types 5 and 6, while combining the middle category with the bottom category (column 6 with column 11) would eliminate the distinction between scale types 1 and 2. In a similar way for question 61, combining columns 3 and 7 would combine scale types 4 and 5, while combining columns 7 and 12 would combine scale types 0 and 1. As has been pointed out, combining answer categories does not affect the rank order of individuals, but it does cut down on the number of ranks between which the scale discriminates. The number of ranks into which the scale continuum can be divided is a direct function of the diversity of marginal frequencies of the items used.

It is worth noting that the combination of answer categories may be quite different from that expected on the basis of manifest content. In both of the present examples of trichotomized items, the following combinations were indicated:

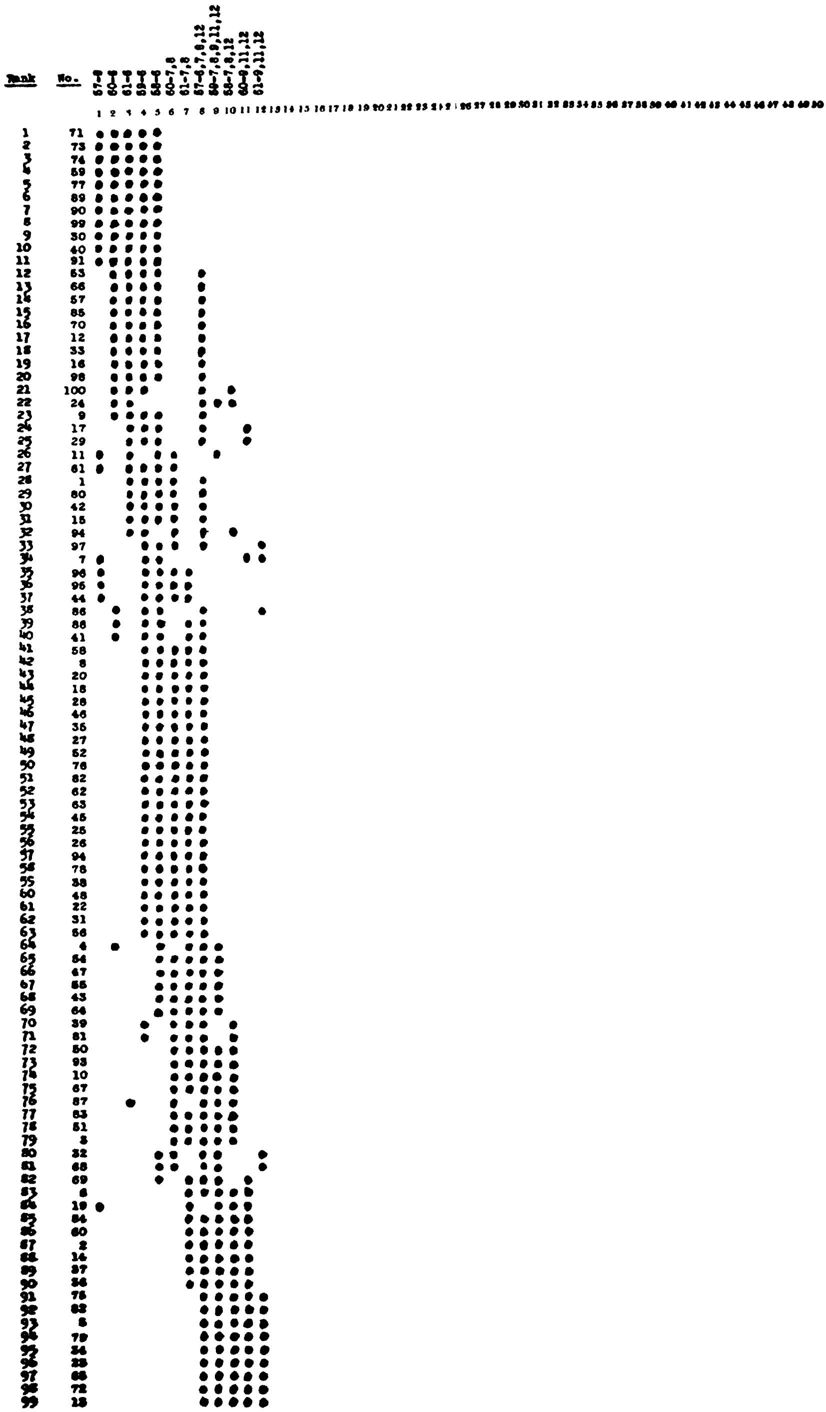
"Positive"	"Strongly agree"
"Neutral"	"Agree" and "Undecided"
"Negative"	"Disagree," "Strongly disagree," and "No answer."

It should be remembered that to a certain degree these combinations are arbitrary, and that where it is desired to base the combinations solely upon the scale pattern, it is advisable to increase the working sample of respondents from 100 to 200 cases, using two boards.

5. *Attitude toward an Army career.* This scale on the desire of enlisted men to remain in the Army after the war, while based on only six dichotomized items, seems almost perfect; the sample shows the extremely high reproducibility coefficient of .98. The six items have marginal frequencies ranging from 12 per cent "favorable" to 59 per cent "favorable." This scale was used in a problem of predicting which men would be most likely to want to stay in the Regular Army after the war. While the prediction of *how many* requires more than just the scale data, the presence of a scale pat-

SCALOGRAM No. 4. ATTITUDE TOWARD POSTWAR CONSCRIPTION (Coefficient of Reproducilibity O-E .94)

Respondent order *Category order*



tern enables the research worker to know that a single variable underlies the series of predictive questions asked. To a certain extent, the high reproducibility of this scale is probably due to the emphasis each of the questions places upon the conditions under which the respondent could be induced to remain in the Army. Since these conditions may be thought of as cumulatively more favorable to reenlistment, the resultant scale pattern is a logical conclusion. In this connection it is worth noting that the single question, "Regardless of what you want to do, do you think you will actually

Scalogram No. 4. Attitude toward Postwar Conscription

Questions and Answer Categories

57. After the war, do you think the United States could or could not have an Army sufficient for the country's needs by taking volunteers only?
    - 6 ______ Yes, I'm almost sure it could
    - 7 ______ Yes, I think it could, but I'm not sure
    - 8 ______ No, it probably could not
    - 9 ______ No, I'm sure it could not

58. After the war, do you think the United States should draft all young men for a certain amount of Army training or should we go back to the regular Army system of volunteers only?
    - 6 ______ Draft all young men for a certain amount of training
    - 7 ______ Go back to taking volunteers only
    - 8 ______ Undecided

59. If you had a son, would you want him to have a certain amount of Army training during peacetime, after this war, or not?
    - 6 ______ Yes, I'm almost sure I would
    - 7 ______ Yes, I think I would, but I'm not sure
    - 8 ______ No, I probably would not
    - 9 ______ No, I'm sure I would not
    - 11 ______ Undecided

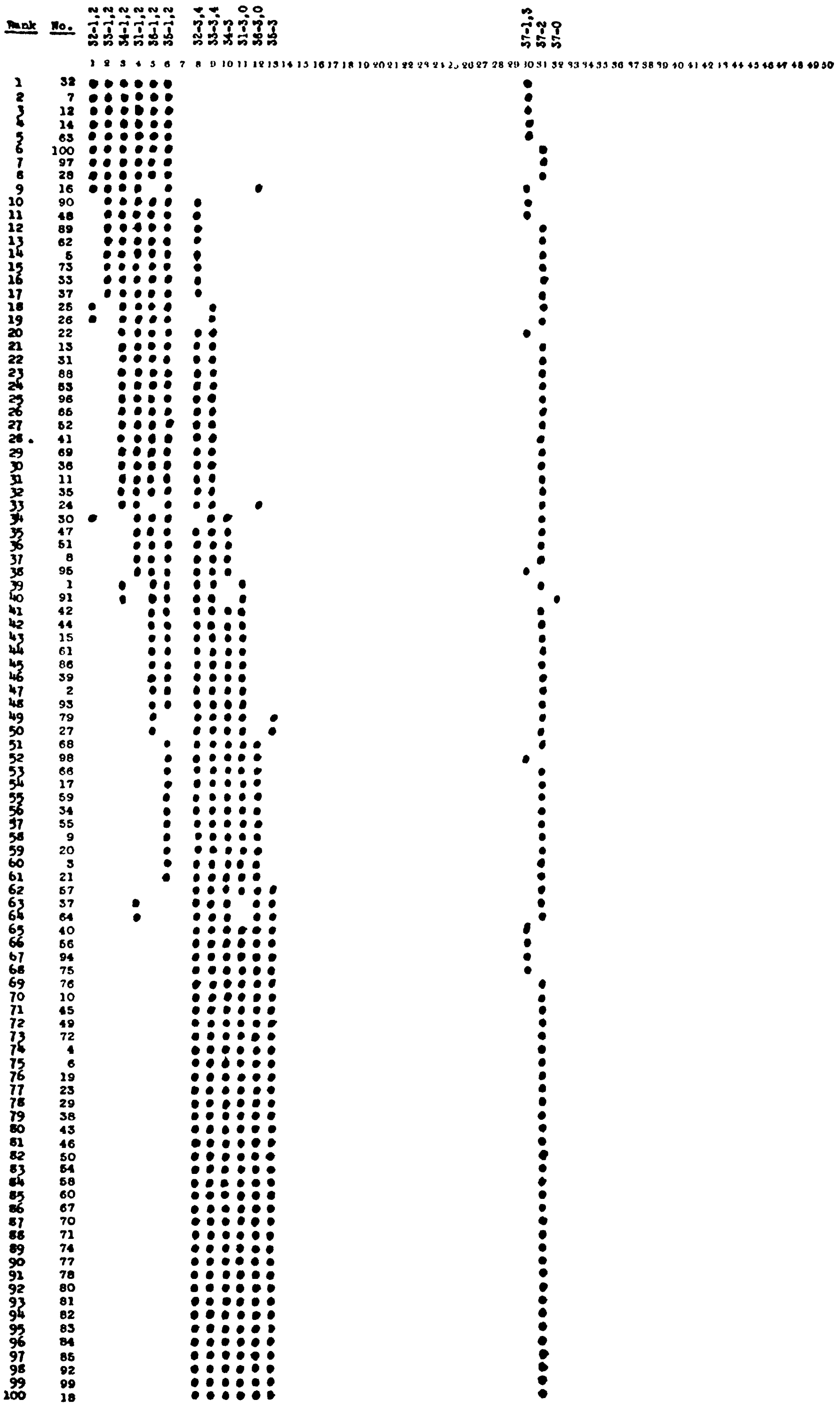
60. Do you agree or disagree with the statement?
    "Even if all other countries agree not to have large armies, the United States should still draft all young men for military training."
    - 6 ______ Strongly agree
    - 7 ______ Agree
    - 8 ______ Undecided
    - 9 ______ Disagree
    - 11 ______ Strongly disagree

61. Do you agree or disagree with the statement?
    "The best way to protect the United States against another war is to make her so strong militarily that no one would dare to attack her."
    - 6 ______ Strongly agree
    - 7 ______ Agree
    - 8 ______ Undecided
    - 9 ______ Disagree
    - 11 ______ Strongly disagree

"No answers" are all coded 12.

SCALOGRAM NO. 5. ATTITUDE TOWARD AN ARMY CAREER (ENLISTED MEN) (Coefficient of Reproducibility = .98)

Respondent order *Category order*



stay in the Army after the war?" fell out of the scale pattern (question 37—column 30, with ten errors and eight nonerrors). This suggests that while the universe being studied is one of desire to stay in the Army (which was the content used in the construc-

Scalogram No. 5. Attitude toward an Army Career (Enlisted Men)

Questions and Answer Categories

31. Are there any conditions you can think of under which you might consider staying in the Army after the war?
    1 ______ I will probably stay in the Army after the war
    2 ______ Under certain conditions I might consider staying in the Army
    3 ______ I *wouldn't* consider it under *any* condition

32. Do you think that you might want to stay in the Army for a *career* after the war?
    1 ______ Yes, I would want to very much
    2 ______ Yes, I might want to, but I'm not sure
    3 ______ No, I don't think I'd want to
    4 ______ No, I would *not* want to at all
    5 ______ Undecided

33. Do you think that you might want to stay in the Army for a *year or two* after the war?
    1 ______ Yes, I would want to very much
    2 ______ Yes, I might want to, but I'm not sure
    3 ______ No, I don't think I'd want to
    4 ______ No, I would *not* want to at all
    5 ______ Undecided

34. If we have a large Army after the war (say 2,000,000 men) do you think you might want to stay?
    1 ______ Yes, I'd like to stay on if we have a large Army
    2 ______ I might want to stay if we have a large Army, but I'm not sure about it
    3 ______ A large Army would make no difference. I wouldn't stay on anyway
    4 ______ Undecided

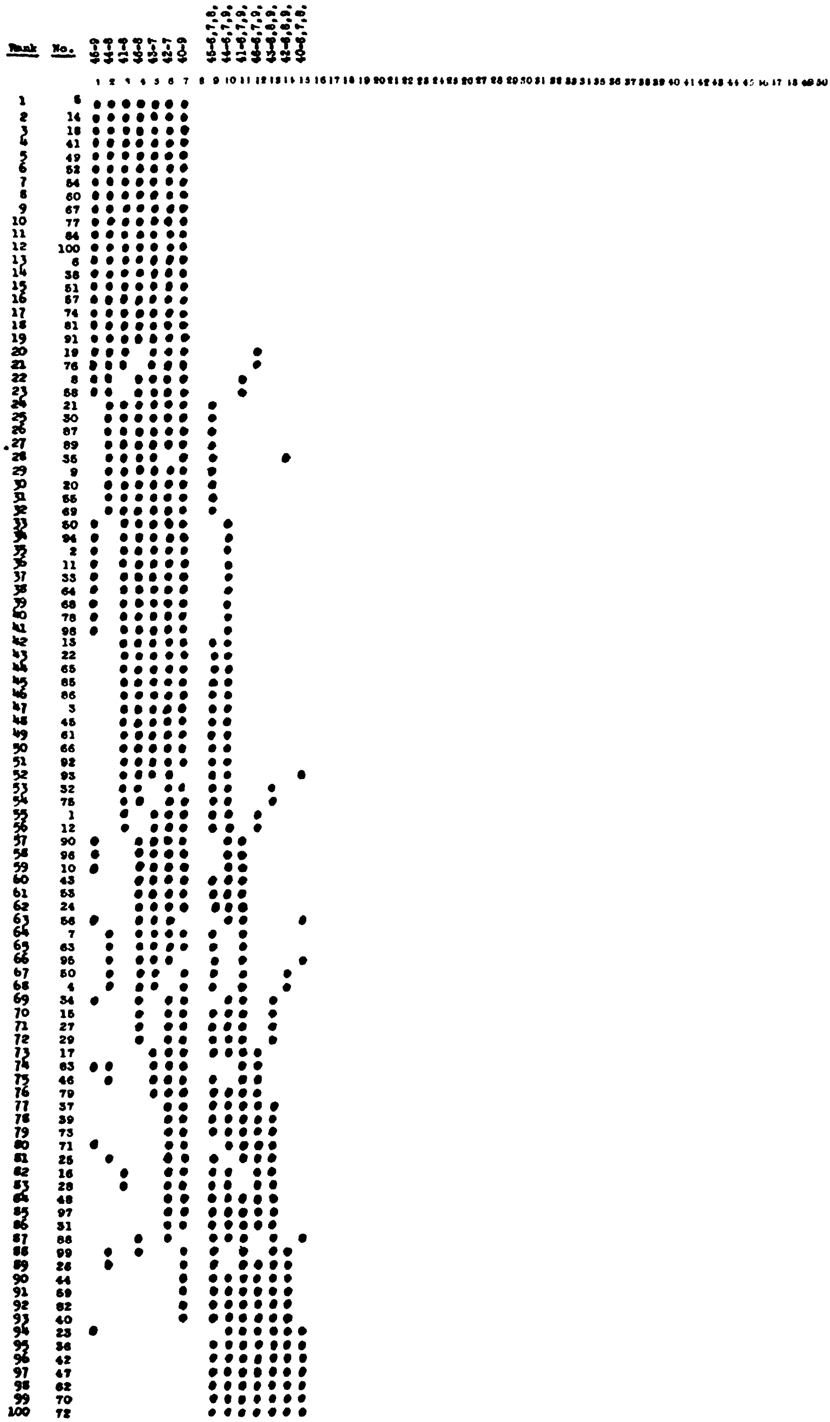
35. If civilian jobs are hard to get after the war, do you think you might want to stay in the Army?
    1 ______ I would want to stay in the Army
    2 ______ I would want to get out to look for a job, but I might want to come back to the Army
    3 ______ I would rather take any kind of job outside than stay in the Army
    4 ______ Undecided

36. Do you think you might want to stay in the Army after the war if you could change your Army job?
    1 ______ I would want to stay, whether I could change my job or not
    2 ______ I might want to stay if I could change my Army job
    3 ______ I wouldn't want to stay, no matter what Army job was offered me
    4 ______ Undecided

37. Regardless of what you want to do, do you think you will actually stay in the Army after the war?
    1 ______ Yes
    2 ______ No
    3 ______ Undecided

"No answers" are all coded 0.

Respondent order *Category order*



tion of the sample of items), there are individuals who *planned* to stay in although they *wanted* to get out, and vice versa. The "desire" to stay in the Army and the "plan" to stay in represent two separate, although related, areas. The present scale represents a ranking of enlisted men according to their desire to stay in the Army; how this desire relates to actual reenlistment is a separate, although extremely interesting, problem. The internal validity of a scale which purports only to *rank* men on a continuum of *desire* to

Scalogram No. 6. Knowledge of Current Events

Questions and Answer Categories

40. Who is the Chief of Staff of the U.S. Army?
    6 ______ General MacArthur
    7 ______ General Eisenhower
    8 ______ General Arnold
    9 ______ General Marshall

41. Which of the following countries is at war against the Axis?
    6 ______ Argentina
    7 ______ Sweden
    8 ______ Brazil
    9 ______ Turkey

42. Where was the first major beach head established by the Allies in their invasion of Europe?
    6 ______ Southern France
    7 ______ Normandy
    8 ______ Holland
    9 ______ Belgium

43. Who is the recently appointed Secretary of State?
    6 ______ Welles
    7 ______ Stettinius
    8 ______ Nelson
    9 ______ Byrnes

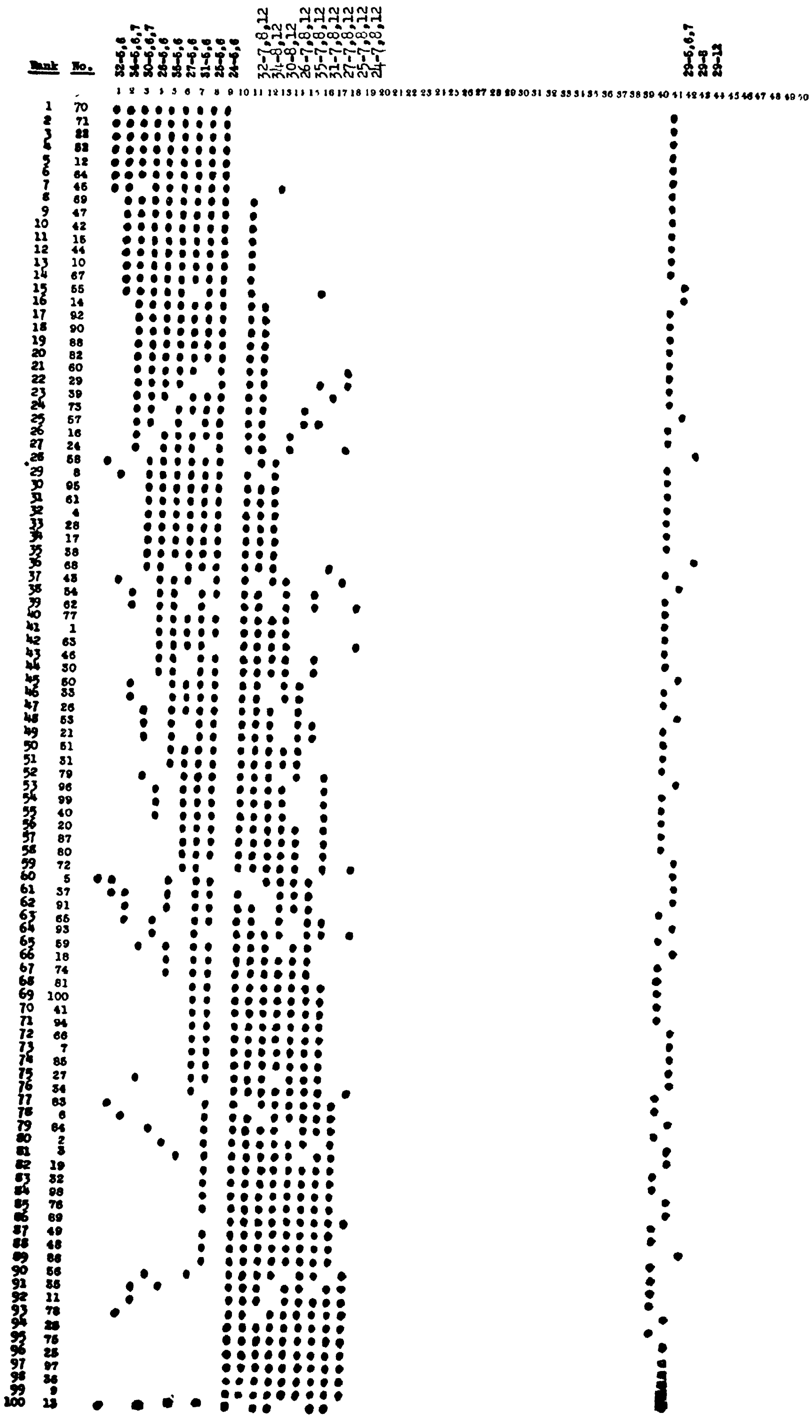
44. The man who said we would have "peace in our time" was
    6 ______ Churchill
    7 ______ Mussolini
    8 ______ Chamberlain
    9 ______ Tanaka

45. The "Co-Prosperity Sphere" was a scheme of
    6 ______ The Germans
    7 ______ The Spanish
    8 ______ The Italians
    9 ______ The Japanese

46. "Fascism" began in
    6 ______ Ireland
    7 ______ Russia
    8 ______ Italy
    9 ______ Germany

"No answers" are all coded 12.

Respondent order *Category order*



do something cannot be judged by an external criterion of behavior which reflects many *other* variables besides the desire.

6. *Knowledge of current events.* Information items can also form scales. When they do, a score on such a scale has a definite meaning. If answers are scored as "right" or "wrong," a person with a higher score must know everything that a person with a lower score knows *and* something additional. Ranking individuals in order of knowledge can be done with the assurance that additional questions of knowledge on the same topic will always produce essentially the same rank order for the people.

The present example is presented for illustrative purposes only. The fact that only seven items were used, each of which was dichotomized, plus the concentration of four of these items at 70 per cent correct or above, does not indicate high reliability.

Scalogram 6 shows the pattern obtained from seven questions dealing with current events. All items are dichotomized into "right" or "wrong," and the coefficient of reproducibility is .93. From this sample we see that, within the limits of error, if a person knows that "the 'Co-Prosperity Sphere' was a scheme of the Japanese," he will know the answers to all the other questions. The scalogram can help one decide whether an information question is really getting at attitudes rather than information—or, stated another way, how much affective tone is involved in what appears to

Scalogram No. 7. Fear Symptoms

Questions and Answer Categories

Soldiers who have been under fire report different *physical reactions to the dangers of battle.* Some of these are given in the following list. How often have you had these reactions when you were under fire? *Check one answer after each of the reactions listed to show how often you had the reaction.* Please do it carefully.

24 ______ Violent pounding of the heart
25 ______ Sinking feeling of the stomach
26 ______ Feeling of weakness or feeling faint
27 ______ Feeling sick at the stomach
29 ______ Cold sweat
30 ______ Vomiting
31 ______ Shaking or trembling all over
32 ______ Urinating in pants
34 ______ Losing control of the bowels
35 ______ Feeling of stiffness

The answer categories for each of the above were:

5 ______ Often
6 ______ Sometimes
7 ______ Once
8 ______ Never
12 ______ No answer

be a factual statement. It is to be expected that if an attitude item were included in this series of information questions, individuals with a "favorable" attitude would occur frequently enough among both the well and badly informed so as not to satisfy the rather rigorous requirements of reproducibility. The same would be true of an information item included in a series of attitude items.

7. *Fear symptoms.* An interesting application of scale analysis is given by the scalogram of the reports of combat veterans on ten different physiological fear symptoms. All but one of the symptoms, "Cold sweat" (columns 41, 42, 43) fit into a scale pattern. In almost all cases the answer categories "Often" and "Sometimes" are combined to signify a "positive" report on the occurrence of the symptoms, while "Once" and "Never" are "negative." Nine dichotomized items result in a coefficient of reproducibility of .92, and the frequencies range from 9 per cent who reported "urinating in pants" to 84 per cent who experienced "violent pounding of heart." The rank order of the dichotomized symptoms permits one to predict, for example, that if a man experienced "shaking or trembling all over" (column 7) he must also have experienced "sinking feeling of the stomach" (column 8) and "violent pounding of the heart" (column 9).

In other words, the symptoms come from a single universe and permit a rank ordering of respondents along a single continuum. There is an intrinsic interdependence among the different fear symptoms which permits them to be ordered from more to less severe. In this case the underlying continuum is probably physiological; in the case of attitudes, the ability to rank items from more to less severe is probably due to the similarity of cultural influences for the population studied.

* * * * *

Many different methodological uses of scalogram analysis have been illustrated in the above discussion of actual scale pictures. Many more examples could be given for each of the applications discussed; other examples could be given to show additional applications. It can be expected that further research will serve to discover even more applications.

The following section will now present a general summary of the utility of the scalogram as a means of scale analysis and point out some of the advantages and ramifications of scale analysis.

SECTION II

THE UTILITY AND RAMIFICATIONS OF SCALOGRAM ANALYSIS

The previous section has presented several examples of scale analyses made by means of the scalogram board. These examples were discussed from two points of view: (1) how did the scalogram board help to determine whether or not a scale existed, and (2) how did the knowledge of whether or not the area being studied was scalable help in the analysis of the problem. The first aspect deals with the research advantages of the visual multicorrelational analysis permitted by the mechanical device called the scalogram board, while the second aspect deals with the much broader problem of the research advantages of knowing that an area is scalable, this is, that the series of items composing the scale permit a rank order of individuals along a single continuum.

It is important to keep clearly in mind the distinction between the advantages of a *test for scalability of items*, such as the scalogram board affords, and the subsequent advantages of the *use of scales* for research, if the test shows that a scale exists. The question, "How does a scale analysis as carried out on the scalogram board (or by the Cornell technique or by some other equivalent procedure) help the research worker in his analysis of a series of items?" is quite different from the question, "How does the fact that a series of items forms a scale help the research worker in the use he makes of such a series of items?" In this section we shall first present an answer to the first question, while an answer to the second question will be given in a later part. The final section of this chapter will present some of the further ramifications which the problem of scaling holds for future attitude research.

Utility of the Scalogram Board

The scalogram board derives its utility from the fact that the basic theory of the present concept of scale analysis calls for a simple, although highly restrictive, pattern of interrelationship between the items being tested. This pattern, as was discussed in detail in Chapter 3, requires that only one possible combination of item responses characterize each scale rank, and furthermore that each scale rank differ from the scale rank above and below it by only one item response. Thus the number of types observed in a scal-

able area is greatly reduced from the total number of possible types. For example, a scalable area of ten dichotomous questions would call for only eleven scale types out of a possible 1,024 types. It is this highly restrictive nature of scalogram analysis which permits the visual analysis of the multivariate frequency distribution of responses of a large number of items for the scale pattern.

The parallelogram pattern described in Chapter 3 is one means of representing the item interrelationships in a scalable area. By arranging individuals in descending order of "favorableness" and by arranging item responses in ascending order of frequency of "favorable" responses, a scalable series of items will result in a parallelogram pattern. If a device could be constructed which permitted both the ordering of individuals and the ordering of items, then it should be possible to test a series of items very quickly for the required parallelogram pattern of a scale. The scalogram boards are just such a device.

The scalogram boards permit the research worker to shift the rank order of individuals while the order of the items remains constant, and vice versa, enabling him to determine visually whether or not the desired parallelogram pattern is forthcoming. Without such a visual aid, the number of multiple correlations that would have to be computed would be staggering. By means of the scalogram board, the research worker can see just how the various items relate to each other and to the pattern as a whole.

The Cornell technique and other equivalent procedures referred to at the beginning of Chapter 4 provide the same picture without using actual boards. Although less flexible, the alternative techniques also result in the visual scalogram desired.

The helpfulness of a scalogram analysis lies in two directions: (1) the type of general pattern that is observed can help in the further definition of the problem, while (2) the specific picture of the individual item and its answer categories can help in the construction of items. Examples of these two uses of scalogram analysis are given below.

1. *Scalogram Analysis and the Definition of the Problem*

It is often found during a preliminary scalogram analysis that a series of questions that were thought to contain only a single variable actually do not scale, but that if the series is broken down into subseries, these subseries do form scales. The researcher can now reexamine his original hypothesis and perhaps reformulate his prob-

lem in terms of multiple variables instead of a single common variable. Once he has redefined his problem in terms of subuniverses of items, he is in a position to construct new items characterizing the new subuniverses and to test his revised hypothesis. It is important to remember that scalogram analysis itself cannot define one's problem; it can only serve to suggest alternative definitions *which then must be tested in their own right.*

A second use of scalogram analysis during the definition stage of a problem is to indicate the existence of nonscale types. Many research problems involve comparisons of different types of groups, groups which differ in *kind,* rather than *degree.* The preliminary scalogram can often show quite clearly when the problem is one of differences in kind rather than degree. Furthermore, it can often point out just what the main types are, their characteristics and relative frequency. On the basis of this information, the researcher may decide whether or not he has to rethink his problem in terms of a typological analysis rather than one involving rank orders.

The following two examples will illustrate the utility of a scalogram analysis for the definition of a problem.

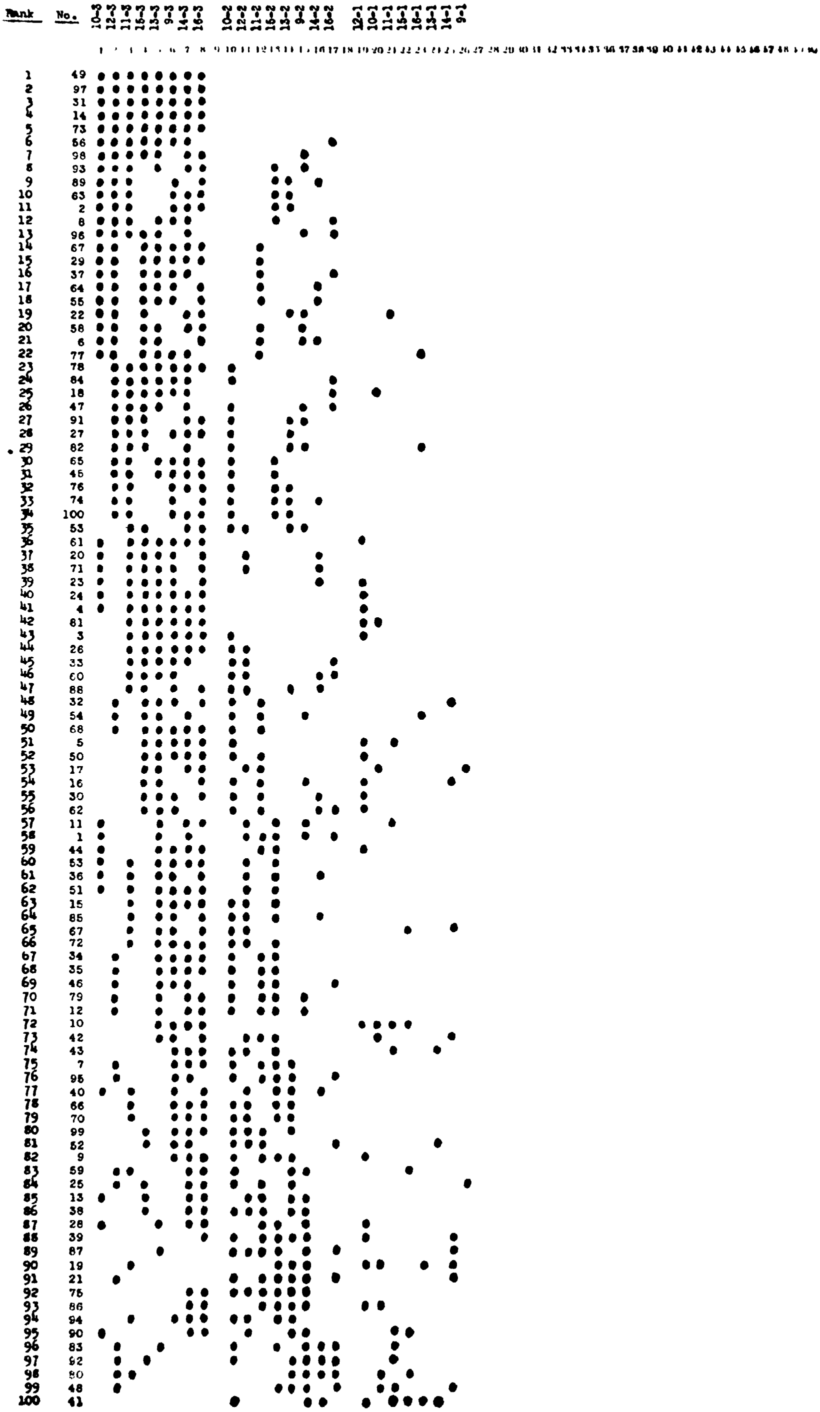
a. *Attitudes toward "sweating out a jump": nonscale types.* Scalogram 8 shows a great deal of error (coefficient of reproducibility = .77); therefore we cannot conclude that there is a single variable in the area of "worries in anticipation of a parachute jump." The differences among individuals appear to be differences in *kind* of worry, rather than *degree* of worry. Soldiers may be characterized by the *kinds* of things they are afraid of in making a parachute jump, rather than the degree to which they are afraid. A further analysis of this problem might take the form of a typological comparison.

Contrast this result with that obtained previously with the universe of fear symptoms among combat soldiers (Scalogram 7). In that example, symptoms were found to scale and we could speak of the degree of fear present. In the present series of items we cannot speak of the relative amount of fear shown by paratroopers, but of the different kinds of fears they exhibited.

b. *Attitude toward combat.* This example, showing how the existence of a nonscale type serves to indicate a possible second variable, will illustrate the way in which scalogram analysis may help to break down a nonscalable area into scalable subareas. A series of six questions dealing with "readiness for combat" appears to form a scale whose coefficient of reproducibility is .94 (Scalogram 9). However,

ducibility = .77)

Respondent order *Category order*



Scalogram No. 8. Attitude toward "Sweating Out a Jump"

Questions and Answer Categories

Different men say that they sweat out different parts of the jump. Check each one of the following list of items to show how much you sweat out each part of the jump.

During the jump today did you sweat out:

9 Whether you would get a rough landing?

1 ______ I sweat out the idea of a rough landing a lot
2 ______ I sweat out the idea of a rough landing some, but not too much
3 ______ I did not sweat out the idea of a rough landing much at all

10. Waiting around in the sweat-shed?

1 ______ I sweat out waiting around a lot
2 ______ I sweat out waiting around some, but not too much
3 ______ I did not sweat out waiting around much at all

11. Whether you would get a hard opening shock?

1 ______ I sweat out the idea of a hard opening shock a lot
2 ______ I sweat out the idea of a hard opening shock some, but not too much
3 ______ I did not sweat out the idea of a hard opening shock much at all

12. Whether you would freeze in the door?

1 ______ I sweat out the idea of freezing in the door a lot
2 ______ I sweat out the idea of freezing in the door some, but not too much
3 ______ I did not sweat out the idea of freezing in the door much at all

13. Whether the chute might fail to open or malfunction?

1 ______ I sweat out the idea of the chute failing to open a lot
2 ______ I sweat out the idea of the chute failing to open some, but not too much
3 ______ I did not sweat out the idea of the chute failing to open much at all

14. Riding in the airplane?

1 ______ I sweat out the ride in the plane a lot
2 ______ I sweat out the ride in the plane some, but not too much
3 ______ I did not sweat out the ride in the plane much at all

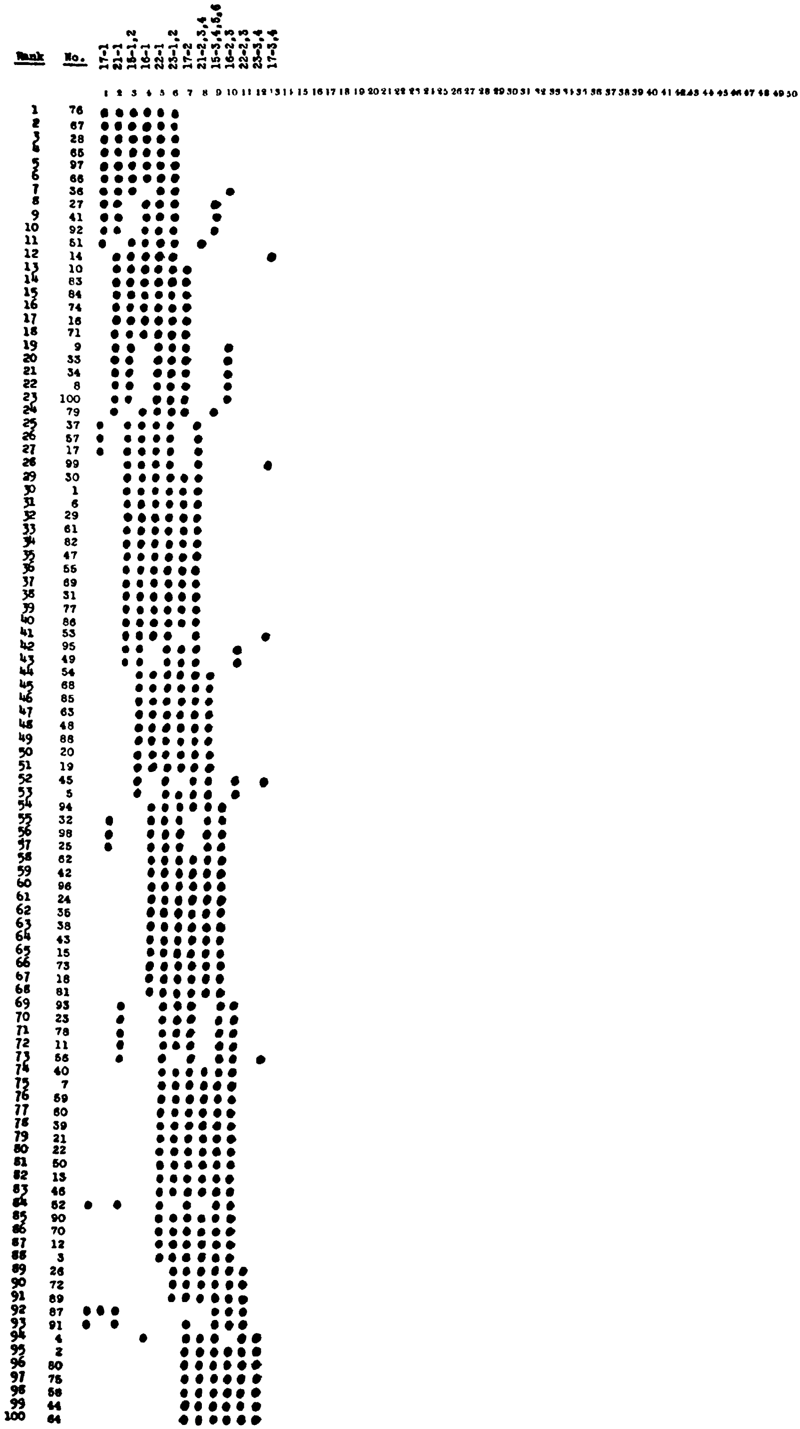
15. The free-fall in the air before the chute opens?

1 ______ I sweat out the idea of the free-fall a lot
2 ______ I sweat out the idea of the free-fall some, but not too much
3 ______ I did not sweat out the idea of the free-fall much at all

16. Whether you might get tangled up with another man on the way down?

1 ______ I sweat out the idea of getting tangled up a lot
2 ______ I sweat out the idea of getting tangled up some, but not too much
3 ______ I did not sweat out the idea of getting tangled up much at all

Respondent order *Category order*



two nonscale types of five cases each appear in ranks 19 to 23 and 69 to 73. The errors producing these nonscale types occur in columns 3 and 10, which contain the answers to questions 15 and 16. A cross tabulation of these two questions alone seems to indicate

Scalogram No. 9. Attitude toward Combat

Questions and Answer Categories

15. Which of the following best tells the way you feel about getting into the actual battle zone?

    1 ______ I want very much to get into it just as soon as possible
    2 ______ I'm ready to go anytime
    3 ______ I'd like to go before it's over, but I don't think I'm ready yet
    4 ______ I hope I won't have to go, but if I do I think I'll do all right
    5 ______ I hope I won't have to go because I don't think I would do very well
    6 ______ No opinion

16. Do you feel that you are now trained and ready for combat or do you need more training?

    1 ______ I'm ready for combat now
    2 ______ I need a little more of some kinds of training
    3 ______ I need a lot more of some kinds of training

17. If and when you get into combat how well do you think you will stand up under battle conditions?

    1 ______ Very well
    2 ______ Fairly well
    3 ______ Not very well
    4 ______ Not well at all

21. Do you think that you are in good physical condition?

    1 ______ I feel that way nearly all of the time
    2 ______ I feel that way fairly often
    3 ______ I feel that way only once in a while
    4 ______ I almost never feel that way

22. Do you think that you are in tough enough physical condition for going into combat?

    1 ______ Yes
    2 ______ No
    3 ______ Undecided

23. Which of the following best describe your own feeling about getting into combat against the Germans?

    1 ______ I'd like to get into the fight as soon as I can
    2 ______ I'm ready to go when my turn comes
    3 ______ I'd just as soon stay out of combat if possible
    4 ______ I don't want to get into combat at all

that more than one variable is involved, as can be seen from the following table:

Question 16: "Do you feel you are now trained and ready for combat or do you need more training?"	*Question 15:* "Which of the following best tells the way you feel about getting into the actual battle zone?" "I want very much to get into it just as soon as possible" and "I'm ready to go anytime"	Other replies	Total
"I'm ready for combat now"	33	14	47
"I need a little more" and "I need a lot more of some kinds of training"	14	39	53
Total	47	53	100

The fact that we have substantial frequencies in *all* four cells indicates the presence of more than one variable. These two questions cannot be scored together if one wishes to rank individuals on readiness for combat. Analysis of the content of these two questions suggests that we have four types of individuals: (a) those who desire combat, and who feel they have had enough training; (b) those who desire combat, but who feel they need more training; (c) those who do *not* desire combat, although they feel they have had enough training; and (d) those who do *not* desire combat and who feel they need more train ng. The hypothesis for further testing would be that the desire for combat and the attitude toward adequacy of training are two separate areas each of which may be scalable separately. To test this hypothesis, one would have to construct additional questions in both areas and test them for scalability.

2. *Scalogram Analysis and Question or Item Wording*

Scalogram analysis provides a powerful test of the internal structure of a series of items. In the field of attitude analysis and public opinion polling, the selection of items and the determination of their relationship to each other is a problem of primary importance. Scalogram analysis offers a complete picture of such interrelationships between items.

In public opinion polling in particular, the problem of the formulation of a "valid" opinion question has revolved around the prob-

lem of the meaning of questions. One of the basic concerns of all public opinion analysts is the interpretation which the respondent places upon the question asked. Rules have been suggested by many people on how to word questions. The pitfalls are well known, but even the experts are often in disagreement about the proper solution. Scalogram analysis does appear to offer some help on this problem. Applied during the pretest stages of an opinion study, scalogram analysis can contribute greatly to the interpretation of the meaning of an opinion question.

Some of the many contributions of scalogram analysis to the problem of question wording are discussed below. They all concern the basic question, "How did the respondent interpret the meaning of this question?" This applies to both the question wording and the answer alternatives. The important theoretical point involved is that any single question asked is but a representative of all other questions that might have been asked instead. To determine the relation a particular question has to the universe of all opinion questions on the same topic, it is necessary to pretest a *series* of questions characterizing the universe. On the basis of the scalogram pattern, one can then select the final question to be used. If it is desired to divide the population into but two or three groups with respect to their attitude, then the one or two questions that have the proper marginal frequencies could be used. If the population is to be divided into an upper half and lower half, then one or two items can suffice to provide this 50–50 split. Or, if it is desired to divide the population into thirds, the one or two items that will yield three ranks with 33⅓ per cent of the population in each are the ones to be selected. Thus, only one or two items can be selected from among all characterizing items to be used in the final study to answer questions like: How does the more favorable half of the population differ from the less favorable half with respect to their age, political party affiliation, etc., etc.? It is quite another matter, it should be clear, to answer the question: What per cent of the population is "favorable?" For this, an objective zero point is needed, a problem which is studied by the intensity function discussed in Chapter 7.

Once a question is selected for the final study, the researcher has the important assurance that this question has a single meaning to all respondents, and that it will rank respondents in the same relative order of "favorableness" as any other questions that one might choose to ask on the same issue.

a. *Problems of interpretation.* Scalogram analysis will help an-

swer such problems as: Did different people interpret the question differently? Was it double-barreled? Was the subject matter too complex or technical to be understood? Did the question contain some unforeseen implication? Was the question wording simple enough to be understood? Did the interviewer change the original meaning of the question? Did the context in which the question was asked affect its meaning? etc. All these problems involve the determination of how well the question fits into the parallelogram pattern, that is, of how many errors of reproducibility there are. If a question turns out to be but a simple function of the rank order on the entire scale, then we know that this question has a dominant single meaning which has not been destroyed by any of the complications listed above. The amount of error gives us an objective test of the singleness of meaning which the questions have for the respondent. To be sure, scalogram analysis will not tell us *why* a question has errors of reproducibility, but it will tell us that there is more than a single interpretation present in the responses to the question.

Any change in wording, or in the way the question is asked, which changes the meaning of the question by introducing a new and different variable or factor, can be detected by its effect upon the scalogram. The introduction of a stereotype or overtone into a question which has the effect of changing the meaning of the question by bringing in some extraneous consideration would be indicated by an increase in the amount of error. If a question changed from a point of opinion to one of information or vice versa, this could also be determined. In this way it is possible to test the effect of changes in question wording or administration upon the meaning which the question has for the respondent.

b. *Problems of answer alternatives.* An inseparable part of the question asked is the form of the answer categories.[2] In this respect we have such problems as: How many alternatives should be given? What order should they take? Should they be treated as separate categories or combined? If combined, which categories should be thrown together? etc. Scalogram analysis will indicate very clearly which answer alternatives have similar meanings and which differ and by how much. Scalogram analysis will show how the answer categories differ in degree. It will show which categories should be combined and what the effect of this combination will be

[2] The discussion is restricted here to the check-list type of answers. Scale analysis, of course, applies to free answers as well, and indeed to any form of collecting data.

upon the ability of the question to discriminate between scale types.

The scale pattern of a question shows how the different answer alternatives differentiate between the rank order of individuals based upon the entire series of questions. If an answer category does not differ in degree from the answer category above or below it, this will be seen by the inability of this category to select a different group of individuals on the rank order from the adjacent categories. Similarly "no answer" and "undecided" categories can be combined with other answer categories on the basis of the general rank order of the respondents who give these answers. Actual illustrations of the use of scalogram analysis in the combination of answer categories were given in the previous chapter. In this way it is possible to pretest the meaning of answer categories and to secure the best combination of alternatives.

Advantages of Working with Scales

The above section has pointed out some of the many ways in which a scalogram analysis can help in the study of a series of items. We have seen how the search for a parallelogram pattern can aid the research worker in the task of defining his problem and constructing his items. Now we come to a more basic question, "Of what use is it to look for scales?" Stated another way, this question could read, "What are the advantages of working with a series of items which can be shown to belong to a scalable universe?"

The answer to be presented in this section deals with two types of advantages, mathematical and conceptual. These advantages spring from the basic fact that a scalable series of items permits the determination of a rank order which is independent of the particular sample of items used. The conceptual advantages will be discussed below in relation to the problems of description and interpretation, while the mathematical advantages will be most obvious in relation to the problem of prediction.

In order to show the important analytical advantages of working with a series of items which form a scale, as opposed to a series of items which may not be scalable, let us look at these three types of problems: (1) problems in description, (2) problems in interpretation, and (3) problems in prediction. Most research studies usually involve at least one of these problems. Analysis by means of scales is helpful in each case.

1. *Problems in description.* If the problem is one of describing an attitude or an opinion, the presence of the scale pattern permits

the following conclusions. A person's score tells what his responses were to each question. Persons with the same score show the same scale characteristics. A person with a higher score is "favorable" on all questions that a person with a lower score answers "favorably," and on at least one more in addition. This rank order, furthermore, exists not only for the given series of questions, but is the same as the rank order that would be obtained with any other series of questions in the same area.

These properties are especially important in the case of a single opinion question analysis.[3] An opinion question *selected from a scale* will be known to divide people into the same relative rank order as any other question in that area. Knowing that a respondent "favored" a statement which few people "favored" enables one to know that he would also favor all other statements in the same area "favored" by a larger number of people.

Thus, one important advantage of a scale lies in the rationale it offers for the definition of a single continuum as discussed in detail in Chapter 3. The concept of a single continuum, while widely used in social research, lacks a clear-cut definition. Our definition of a single continuum as a series of items each of which is a simple function of the scale scores permits a clear-cut statement of what is meant by a rank order based on a single variable. The properties described above are a result of working with unidimensional universes thus defined. While it is possible to get a rank order in nonscalable areas by assigning arbitrary weights to the items, this rank order does not possess the important invariant properties of a rank order where the problem of how much to weight each item does not come up.

2. *Problems in interpretation.* Knowing that questions of similar content scale together shows them to be measuring a single attitude or opinion variable and warns one against attempting to interpret them as measuring different (albeit related) variables on the basis of their manifest content. This warning is particularly appropriate

---

[3] For example, Cantril concludes in a chapter on "The Use of Breakdowns": "For most members of the population, any refined attitude scale or rank-ordering is much too complicated to use in the usual interviewing situation. . . . If several relatively simple questions can be devised to tap the same variable, then the answers of a single individual on these several questions can be pooled to place him on the scale with respect to the total population." Scale analysis satisfies perfectly the condition of the latter sentence—several relatively simple questions can be used and tested for scalability.

Hadley Cantril (editor), *Gauging Public Opinion* (Princeton University Press, Princeton, N.J., 1944), p. 191.

when one is tempted to infer some "causal" connection from a cross tabulation. The public opinion literature is full of such "causal" explanations. Two opinion questions are cross tabulated to show, for example, that individuals who do not want OPA continued are among those more likely to object to government subsidies to keep prices down. A hypothetical cross tabulation might look as follows:

Question: Do you approve or disapprove of the government paying money to farmers in order to keep prices down?	*Question:* Do you approve or disapprove of the continuation of OPA?		
	Approve	Disapprove	Total
Approve	50	0	50
Disapprove	30	20	50
Total	80	20	100

An inference that might be made from this table would be that approval of a policy of subsidies "leads" one to approve also of the OPA. Theoretically one can never infer "cause" from the cross tabulation of two attitudes or opinions gathered at the same time.[4] However, opinion analysts make this inference quite often. Doubtful as such an inference is at all times, it is even more doubtful when the scalogram pattern showing a zero frequency in the positive-negative cell is obtained.[5] Such a pattern strongly indicates the presence of but a single variable, i.e., attitude toward governmental control, with the possible interpretation that both questions are simply different aspects of the same one thing. The interpretation of such a pattern can best be made in terms of rankings on a single opinion continuum, rather than in terms of a relationship between two different variables.

3. *Problems in prediction.* If an area is scalable, then it can be used very simply as either the criterion or predictor in a prediction problem.

If the scale items are to be predicted from outside variables, the researcher knows that any result based upon the cross tabulation of one of the scale questions with some other variable (e.g., a background characteristic such as age or sex) will be essentially the same as the result obtained from any of the other questions in the series. It is therefore often unnecessary in making "breakdowns" to tabu-

[4] A good discussion of this problem is given in E. Greenwood, *Experimental Sociology* (King's Crown Press, New York, 1945).

[5] See Chapter 3 for a complete description of this pattern.

late more than one of the questions in the series. For example, if a series of questions is known to scale, then *all* questions in the scale will show men more "favorable" than women, if such is the case. Of course, the most sensitive differentiation will be obtained by use of the scale scores themselves. The single scale question which will show the highest relationship will be that question whose marginal frequency most closely approximates the marginal frequency of the outside variable.

If an outside variable is to be predicted from the scale items, the multiple correlation on the scalable area is equivalent to the zero-order correlation with the scale score. Only the simple correlation with the scale scores need be computed, yet the full predictive power of the area is realized thereby. The multiple correlation must always be equal to the simple correlation of the outside variable with the scale ordering of persons.

The importance of this property for the problem of prediction weights is obvious. In a scalable series of items, the same prediction weights (namely, the scale weights) may be given to the items regardless of what the outside variable may be. This is not the case for a nonscalable area. If a series of items do not scale, then a new set of weights must be computed for each new prediction problem. In this way scale scores provide a quantification of the attributes or responses that is invariably most efficient for predicting any outside variable whatsoever.

Furthermore, it is known that adding more items or questions from the same area will *not* increase the multiple correlation. The maximum predictability in the area can be determined simply by means of the selected sample of items or questions. Since much of public opinion analysis implicitly involves predictions on the basis of only one or two questions in an area, this feature of scale analysis is of utmost importance. One can assign simple weights to a small number of questions from the scalable area with the assurance that the prediction result would be the same as that based upon all the possible questions from the same area. Adding more questions to a sample from a scalable universe will not increase predictive power substantially, so that relatively few items are highly efficient in practice for retaining all the predictive power the universe has for the criterion.

Scale Analysis and the Measurement of Intensity

Finally, scale analysis provides the basis for the further study of a component present in scalable areas—*intensity of feeling*. With

the rank order of individuals along a scale continuum from more to less "favorable," there is implied a related rank order of individuals along an intensity continuum from more to less "*intense*." The measurement of this intensity component and its application to attitude and opinion analysis will be discussed in detail in Chapter 7.

To summarize the utility of scalogram analysis, we find that this utility springs on the one hand from the use of the scalogram analysis for the detection of scales, and on the other hand from the use of scales in the analysis of one's data. In relation to the detection of scales, scalogram analysis helps to define one's problems in terms of smaller subareas or in terms of typological analyses. Scalogram analysis also aids in the construction of items by enabling one to test these items for meaning or "bias" and by indicating how response categories can be combined. In relation to the advantages of working with scales, we find that problems of definition, interpretation, and prediction are greatly simplified if the items can be shown to come from a scalable universe.

SECTION III

FURTHER RAMIFICATIONS OF SCALE ANALYSIS

Work on the present theory of scale analysis has been limited almost entirely to the field of attitudes and opinions. While scale analysis is formal and applies to any type of social science data, the empirical applicability of scale analysis to these other groups of data remains to be investigated.[6] It may be helpful for future research to report some of the experiences which have been encountered in studying scales in the field of attitude and opinion research.

The Occurrence of Scales

A question which is often asked is, "How often do you find scales in practice?" Quite obviously, if the rigid parallelogram pattern required of a scale did not occur empirically, then the theory would have very little practicality. There is a real question, then, as to whether scales occur frequently enough to be applicable to the study of social attitudes.

As has been indicated earlier, the bulk of social phenomena is too complex for one to expect many aspects to be scalable. It would

[6] Since this was written there has appeared the interesting example of scalability of institutionalized discrimination. See Gilbert Shapiro, "Myrdal's Definitions of the 'South': A Methodological Note," *American Sociological Review*, Vol. 13, No. 5 (October 1948), pp. 619–621.

be interesting if we could, on the basis of the experience of the Research Branch, estimate to what extent and in what kinds of data scalability occurs. Unfortunately such an estimate is not warranted at the present time. The work of the Research Branch was of such a nature due to the immediate problems of wartime research that, in many cases, where a series of items proved to be nonscalable, no further attempt was made to see if the conceptual universe could be redefined into subuniverses which might prove to be scalable. The thorough application of the present theory of scale analysis would call for many pretests and revisions of both the content universe and the sample of items. For example, if a small proportion of items from a sample does not conform to the scale pattern that seems to be indicated by the rest, the discrepancies would have to be studied further, perhaps by detailed interviews, to see what the disturbances might be due to. In some cases, it may be found that a subuniverse is indicated that must be treated apart from the rest of the universe; in other cases it may be found that there have been mistakes in wording or presentation of the questions. Thus many cases which were classified as nonscalable upon first examination may have proved, upon additional research, to be scalable, or to be divisible into scalable subuniverses.

However, while it was a much more frequent experience *not* to find a series of items scalable according to the four criteria set forth, there were certainly enough instances of scalable areas to warrant further research with the present theory of scale analysis. It might be helpful for future research to list some of the areas which *upon first approximation* did not prove scalable. Several of these nonscalable areas were attitude toward "sweating out a jump" (presented in the previous section), attitude toward the Russians, attitude toward internationalism (postwar foreign policy), attitude toward obeying Army rules and regulations, attitude toward the Army Score Card Plan (later broken down into two scalable areas, attitude toward the *idea* of the Score Card Plan and attitude toward the *administration* of the Score Card Plan—see Chapter 7 for details), attitude toward future results of the war, attitude toward mail censorship, attitude toward veteran preference for government jobs, attitude toward civilian support of the war effort, attitude toward the Army's concern with personal welfare, attitude toward use of athletes and entertainers in the Army (later broken down into two scalable areas—attitude toward athletes and attitude toward entertainers). Examples of several scalable areas were given in detail in the previous chapter.

A promising avenue for future research might consist of a classification and examination of those areas in social science which prove scalable or nonscalable. It would be particularly helpful to have some systematic analysis of how certain areas, and even individual items which proved nonscalable upon first approximations, were later modified to form scalable areas or items. The present data must be taken almost entirely as *illustrative* of the present theory of scale analysis—especially in relation to the intensity component to be presented in Chapter 7. A great deal of work has yet to be done upon the empirical occurrence of scales.

It is quite obvious from the results obtained by the Research Branch that many attitudinal areas are so complex as to be nonscalable. The possible existence of other systems of order than that of the perfect scale is a problem which must be studied further. Another type of structure which permits of a rank ordering of the people—albeit with a different meaning from that of the perfect scale—is that called a *quasi scale,* as is discussed next.

Quasi Scales

It may happen that a series of items appears to measure a strong common variable, but that there are too many errors of reproducibility to permit adequate reproducibility of responses from scale scores. The errors of reproducibility may be caused either (a) by one or two other important variables that may be in the area, or (b) by many small variables. The existence of one or two additional variables in the area is indicated by *nonscale types* in the scale pattern. If such definite nonscale types exist, then the multiple correlation of an outside variable with the whole area would *not* be quite equivalent to the simple correlation with rank order on the dominant variable, and would be attained only by taking the nonscale types into account. On the other hand, if errors of reproducibility are random, then the multiple correlation of any outside variable on the area will be precisely equal to the simple correlation with the rank order on the area. This property, it is important to note, holds *no matter how low the reproducibility is.*

Some areas which are not scalable in terms of reproducibility are called *quasi scales*; their reproducibility may not be high but their errors occur in a sort of gradient. This gradient pattern of errors indicates that, while there is not a single factor operating as in the case of a scale, nevertheless there is a single dominant factor and *indefinitely many* small random factors, so that prediction of any external variable must rest essentially on the dominant factor. The

dominant factor is measured by the quasi-scale scores. This means that although quasi scales lack an essential property of a scale-rank order, i.e., they cannot reproduce the respondent's characteristics on the items in the area very well—nevertheless, the rank order is perfectly efficient for relating any outside variable to the area. Therefore, if examination of the errors of reproducibility shows them to conform to a certain gradient pattern, and not to be grouped together to form nonscale types, then we have what may be called a quasi scale. The pattern of responses indicates that there is a single dominant factor measured by the quasi-scale scores.

The diagram on the next page will serve to illustrate the difference between random scale errors, grouped nonscale errors, and gradient quasi-scale errors.

The error pattern of the *"true" scale* question is recognizable from the random nature of the few errors of reproducibility that do occur. The cutting point between "positive" and "negative" categories would fall between respondent 18 and 19. Predicting all respondents from 1 to 18 to have given "positive" answers would result in two errors (respondents number 11 and 16); predicting all respondents from 19 to 50 to have given "negative" answers would result in three errors (respondents number 22, 25, and 31). Reproducibility of this item would therefore be high—5 errors out of 50 predictions.

The error pattern of the *nonscale* question is recognizable from the way in which the errors are grouped together. Again locating the cutting point between "positive" and "negative" categories between respondents 18 and 19, we find that predicting all respondents from 1 to 18 to have "positive" answers would result in five errors, and furthermore *that these five errors were grouped together among respondents 8, 9, 10, 11, and 12.* Similarly predicting respondents 19 to 50 to have given "negative" answers, also results in a grouping of six errors among respondents 39, 40, 41, 42, 43, and 44. Reproducibility of this item would therefore be low—11 errors out of 50 predictions—and, more important, these errors would be grouped together to form nonscale types. The grouping of errors into nonscale types indicates that more than one strong variable is present. We cannot find any single rank order of respondents that would successfully represent the attitudes of respondents to both variables.

The error pattern of the *quasi-scale* question is recognizable from the manner in which the fairly large number of errors that occur gradually decrease in number as one moves further and further

away from the cutting point. These errors are much more frequent than "true" scale errors, but do not group together like nonscale errors. Again locating the cutting point between respondents 18

	RANDOM SCALE ERRORS		GROUPED NONSCALE ERRORS		GRADIENT QUASI-SCALE ERRORS	
Respondent	*Categories*		*Categories*		*Categories*	
Rank Order	+	−	+	−	+	−
1	X		X		X	
2	X		X		X	
3	X		X		X	
4	X		X		X	
5	X		X			X
6	X		X		X	
7	X		X		X	
8	X			X		X
9	X			X		X
10	X			X	X	
11		X		X	X	
12	X			X	X	
13	X		X			X
14	X		X			X
15	X		X			X
16		X	X		X	
17	X		X		X	
18	X		X		X	
19		X		X		X
20		X		X		X
21		X		X		X
22	X			X		X
23		X		X		X
24		X		X		X
25	X			X		X
26		X		X		X
27		X		X		X
28		X		X		X
29		X		X		X
30		X		X	X	
31	X			X	X	
32		X		X	X	
33		X		X		X
34		X		X		X
35		X		X		X
36		X		X		X
37		X		X	X	
38		X		X	X	
39		X	X			X
40		X	X			X
41		X	X			X
42		X	X		X	
43		X	X			X
44		X	X			X
45		X		X		X
46		X		X		X
47		X		X		X
48		X		X	X	
49		X		X		X
50		X		X		X

and 19, we find that predicting "positive" responses for respondents 1 to 18 would lead to six errors, and that these errors would occur in groups of three (respondents 13, 14, and 15), then two (respondents 8 and 9) and finally one (respondent 5) in a sort of gradient as one moved away from the cutting point. A similar gradient of errors would occur in the "negative" response predictions for respondents 19 to 50; three errors among respondents 30, 31, and 32, two errors among respondents 37 and 38, one error for respondents 42 and 48 respectively. Reproducibility of this item would be low—13 errors out of 50 predictions—but more important, these errors would not be grouped together into nonscale types, but would consist of small groups of errors gradually decreasing in size as one moved away from the cutting point.

The Research Branch has found evidence of the quasi-scale pattern in the case of personality inventories, information tests, and measures of intensity of feeling. An example of such a quasi scale is given by the scalogram on page 164 based upon an aggregate of fifteen items symptomatic of psychoneurotic tendencies (Scalogram 10). Most items have a large number of errors, although there is enough nonerror to permit the appearance of a fairly definite parallelogram pattern. The coefficient of reproducibility is .73 (this could be raised somewhat by dichotomizing the response categories). In general, the errors for each category conform to a gradient, errors decreasing gradually as the ranks depart from the region of highest density of response for the category.

The importance of a quasi scale lies in how it is used for external prediction problems. While we cannot derive a person's responses from his quasi-scale score, the score does yield a zero-order correlation with any outside variable which is equivalent to the multiple correlation on all the items in the quasi scale. The prediction of the external variable rests essentially on the dominant factor that is being measured by the quasi-scale scores. Thus a quasi scale has the full mathematical advantages of a scalable area.

In the case of a quasi scale, it is ordinarily necessary to use a much larger sample of items than in a scale in order to obtain the proper rank order of individuals.

What implications does the quasi scale hold for scale measurement? Quasi scales seem to enable one to take a large number of items which have some strong common content and to derive from their intercorrelations a score which permits a rank order independent of item weights. It would thus appear that quasi scales

offer a promising avenue of research into some complex areas which are neither scalable nor divisible into scalable subareas. While the single dominant variable of a quasi scale cannot be represented by means of a small number of items due to the amount of error involved, increasing the number of items which contain this dominant variable makes this error assume a gradient pattern, and permits an invariant rank order.

One interesting way in which quasi scales can be seen to arise is as follows. Suppose an attitude universe can be broken down into many subuniverses, each of which is scalable. Suppose further, that when the scores on the subuniverses (which can now be treated as quantitative variables) are intercorrelated, they are found to correspond to Spearman's single common factor structure. Then it must follow that if an item is selected from each of the subuniverses, and a scalogram formed of these items, the scalogram will reveal a quasi scale; and furthermore, the quasi-scale score will be essentially the score on the Spearman common factor.

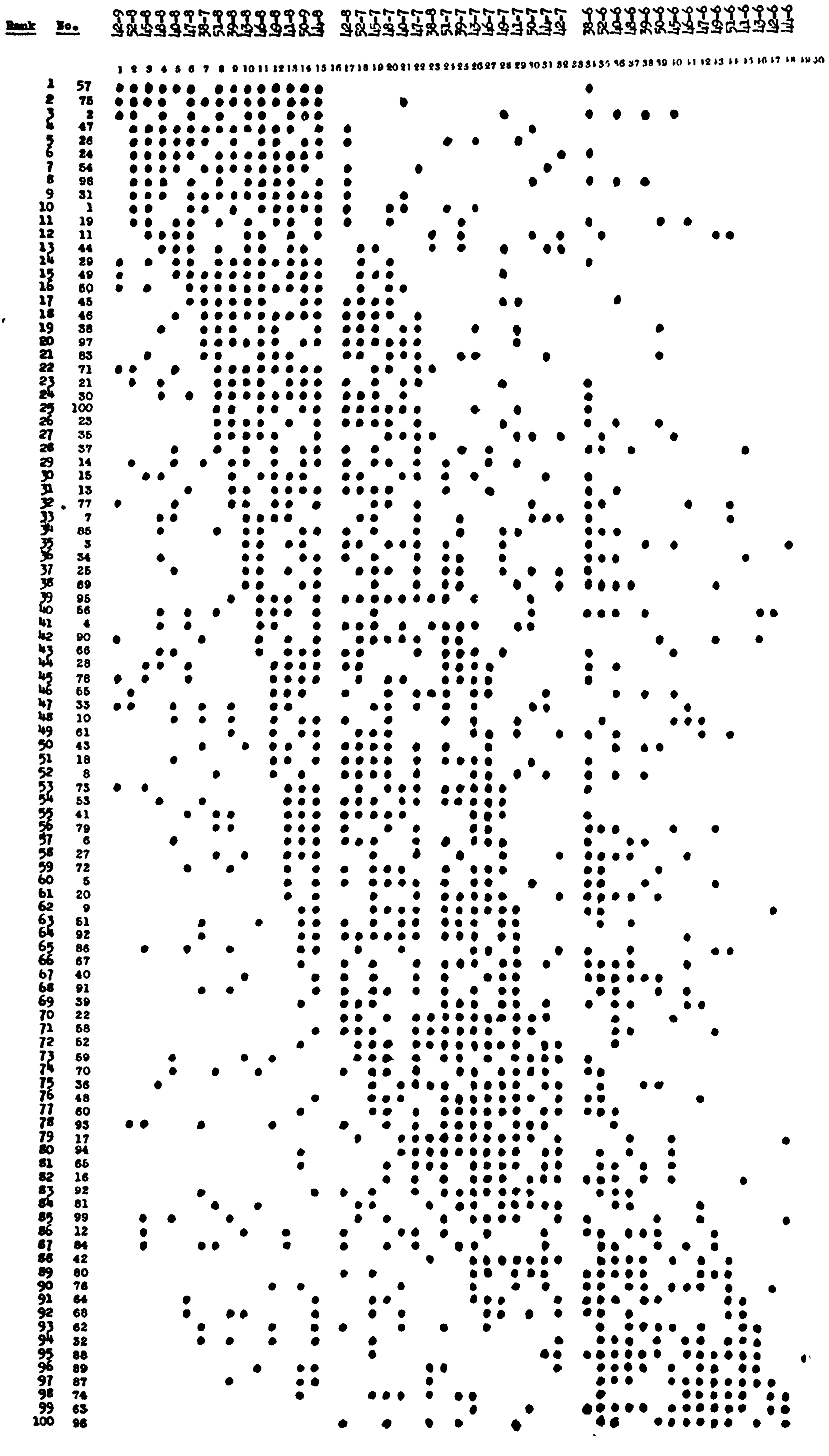
The Research Branch did have a nonscalable universe—morale—for which it found several subuniverses to be scalable, such as pride in one's outfit, satisfaction with one's job, confidence in one's leaders, etc. However, the intercorrelations of scale scores were found not to conform to the Spearman structure, but rather to indicate more than one common factor. This explains why morale was not found to be even a quasi scale, and shows that no single ordering of men is possible for morale as a whole.

Sampling the Universe of Items

A basic assumption of the proposed theory of scale analysis is that a sample of items which were found to be scalable can adequately represent the unlimited universe of similar items. An important property of a scalable universe is that the ordering of persons based on a sample of items will be essentially that based on the universe. Theoretically, then, if a series of ten items were found to be scalable, then all items of the same content would also be scalable.

The process whereby the initial ten items are constructed is an intuitive one. One turns over in one's mind many possible items which characterize the universe in which one is interested. Ten of these are selected as a sample of all possible items that might occur to the research worker, and these ten are then subjected to a scale analysis. If these ten items scale, the assumption is made that all

Respondent order *Category order*



Scalogram No. 10. Psychoneurotic Inventory

Questions and Answer Categories

38. Do you have any particular physical or health problem?
    6 ______ Yes
    7 ______ No
    8 ______ Undecided

39. Have you ever been bothered by shortness of breath when you were not exercising or working hard?
    6 ______ Yes, often
    7 ______ Yes, sometimes
    8 ______ No, never

40. Are you ever troubled by your hands sweating so that they feel damp and clammy?
    6 ______ Yes, often
    7 ______ Yes, sometimes
    8 ______ No, never

41. Are you ever bothered by having nightmares (dreams that frighten or upset you very much)?
    6 ______ Yes, many times
    7 ______ Yes, a few times
    8 ______ No, never

42. How often are you bothered by having an upset stomach?
    6 ______ Nearly all the time
    7 ______ Pretty often
    8 ______ Not very often
    9 ______ Never

43. Have you ever been troubled by "cold sweats"?
    6 ______ Yes, often
    7 ______ Yes, a few times
    8 ______ No, never

44. Have you ever had any fainting spells?
    6 ______ Yes, several times
    7 ______ Yes, a few times
    8 ______ Never had any

45. Have you ever been bothered by your heart beating hard?
    6 ______ Yes, often
    7 ______ Yes, a few times
    8 ______ No, never

46. Have you ever been bothered by pressure or pains in the head?
    6 ______ Yes, often
    7 ______ Yes, sometimes
    8 ______ No, never

47. Have you ever had spells of dizziness?
    6 ______ Yes, many times
    7 ______ Yes, a few times
    8 ______ No, never

48. Do you often have trouble in getting to sleep or staying asleep?
    6 ______ Very often
    7 ______ Sometimes
    8 ______ Almost never

49. Do your hands ever tremble enough to bother you?

6 ______ Yes, often
7 ______ Yes, sometimes
8 ______ No, never

50. Do you ever bite your fingernails now?

6 ______ Yes, often
7 ______ Yes, sometimes
9 ______ No, never

51. Are you ever troubled by sick headaches?

6 ______ Yes, often
7 ______ Yes, sometimes
8 ______ No, never

52. Are you ever bothered by nervousness?

6 ______ Yes, often
7 ______ Yes, sometimes
8 ______ No, never

the items that come to one's mind on the same topic would also scale. If these ten items do *not* scale, by the same token the assumption is made that all the items also would not scale. This is an important assumption, since upon its truthfulness rests the whole concept of inferences about the universe based upon a sample from that universe. How can such an assumption be subjected to empirical verification?

Obviously no fixed universe exists from which one could select random samples and test their representativeness. However, the following suggestion is made for the construction of an "operational" universe which would serve as an important test of the assumption of sampling representativeness. A universe is named and then the research worker compiles a series of perhaps 300 items which characterizes that universe. These 300 items then can be thought of as an "operational" universe. The assumption of sampling could then be tested by selecting at random subseries of 10 items each and testing these for scalability. How many of these subseries scale or do not scale? This result could be compared with the scalability of the entire aggregate of 300 items. In this way, the assumption of representativeness could be tested. The assumption of the invariance of the rank order of scale types could also be tested by means of the correlations between subseries and the total series for scalable areas. See Chapter 8 for a more detailed discussion of this problem, and also for numerical examples of comparisons between subsamples of items.

The Determination of Characterizing Items

The above test could also be expanded to include an experiment in the determination of characterizing items. At present the definition of content is entirely subjective. There is no way to decide whether or not an item belongs to the universe, except through a decision on the part of the investigator or a group of judges. Thus it is possible for one investigator studying an area, such as attitude toward war, to construct a series of items, to test these items for scalability, and to find that they are nonscalable, whereas another investigator studying the same area by means of a different series of items might find them scalable. The answer of course would have to be that since the content of the items themselves defines the area, these two investigators were studying different areas despite the fact that the areas bear the same conceptual title.

The unsatisfactory nature of this answer is obvious. Some less subjective method for defining the characterizing item would help in the selection of sample series. Additional research is greatly needed on this problem. One possible approach might include the use of a group of judges to select the characterizing items whose scalability is to be tested. A combination of the Thurstone technique of evaluating the content of items and the present technique of scale analysis might prove very instructive.[7] While it is probably impossible to remove the element of subjectivity from the process of constructing characterizing items, it should be possible to decrease the danger of individual bias by means of groups of judges.

The Halo Effect

An interesting ramification of scale analysis which deserves further study is the relation between scale analysis and the psychological phenomenon observed in the field of rating "scales" called the "halo" effect. Enlisted men who are asked to evaluate their officers on a series of eleven questions dealing with apparently different aspects of leadership appear to "carry-over" a generally "favorable" or "unfavorable" attitude, instead of answering each

---

[7] An example of this approach is given by A. L. Edwards and F. P. Kilpatrick, "A Technique for the Construction of Attitude Scales," *Journal of Applied Psychology*, Vol. 32, No. 4 (August 1948), pp. 374–384.

question as though it dealt with a different aspect of the officer's behavior (see Scalogram 2 in the first section). It has often been found that questions asking for the evaluation of different groups of people on different traits, including rating schemes for single individuals, are usually so closely related that the researcher is forced to conclude that his subjects are rating some general trait rather than the specific traits inherent in the manifest content of the items in his rating scheme.[8] This phenomenon has often been called a "halo" effect. How does this "halo" effect relate to the present theory of scale analysis?

Scale analysis appears to afford a method for studying "halo" effect in the form of a test of the hypothesis that there is some single variable upon which the different evaluations are based. It carries the "halo" concept into the entire field of attitude or opinion research and shows how only a single variable may underlie the responses to what the analyst believes to be a series of opinion questions on quite different aspects of a problem. And just as the interpretation of separate trait names must be treated with caution once the existence of "halo" has been found, so it would seem the individual interpretation of separate opinion questions must be treated with caution once they have been found to belong to a single scalable universe. "Halo" in this sense may be studied by a test for single meaning. Scale analysis thus may be able to provide an interesting rationale of how individuals can be ranked upon a single rating continuum based upon a sample of items.

The "Relativity" of Scales

Another important ramification of scale analysis concerns the application of scales to different populations or at different periods in time. Scales are relative both to time and to populations. A series of items or questions may form a scale for one population and yet fail to form a scale for another population. Or a series of questions may fail to form a scale for a total population, and yet scale for a subgroup of that population. It would therefore be possible that a nonscalable series of items could be found to form a scale when limited to a subgroup of the population. If such were the case, one could make an analysis of the structure of a content area in terms

[8] As stated by Thorndike, "Ratings were apparently affected by a marked tendency to think of the person in general as rather good or rather inferior and to color the judgments of his qualities by this general feeling." Edward L. Thorndike, "Constant Error in Psychological Ratings," *Journal of Applied Psychology*, Vol. 4, No. 1 (March 1920), p. 25.

of the different meanings involved for different subpopulations. A useful technique for the determination of "deviant" subgroups would be to sort out those individuals who were responsible for scale errors and to study their composition to see if they came from any particular subgroup. Scales based upon one population cannot be transferred to the study of. some other population without first testing to see if the items still scale for the new population. There may be a great fallacy in many cases of applying scales which have been "standardized" on one population to the study of new populations.

One suggested hypothesis for the relativity of scales is that the difference between the two populations is not one of *degree*, but of *kind*. One group is not simply more or less favorable than the other group, but actually defines the issue in a completely different way. The items do not characterize the same universe to the different populations.

When this is the case, when a series of questions scales for one population but not for another, then comparisons among the populations involving statements that one group is "more" or "less" favorable on the same universe than another group must be made with great caution.[9] This is an additional reason for studying attitudes and opinions by means of a series of questions rather than only a single question. In comparisons between groups based upon a single question, one cannot be certain that the different groups attach the same interpretation to the question asked. One cannot predict on the basis of a single question analysis that, if another question were asked on what was conceived to be the same issue, the rank order of "favorableness" among different groups might not change. For example, even if a single question were used only to compare different educational levels on degree of "favorableness" toward an issue, it would first have to be shown that the question had the same meaning, i.e., belonged to the same scalable universe for the different educational groupings, before any comparisons of degree of favorableness could be made.

The above discussion of the relativity of scales according to the population studied applies equally well to the relativity of scales for

[9] The test for scalability requires that the same question be asked of the different populations and that an attempt be made to combine answer categories in the same way, even though these combinations produce slightly more error than they would if each population were scaled separately. If a question scales for two populations, but with different combinations of answer categories, the conclusion would be that the question means the same thing to both populations, but that the answer categories do not represent the same degree of "favorableness."

the same population when two different time periods are studied. A series of items may form a scale at one time for a given population and yet fail to form a scale at another time. Again, the suggested explanation is that the change in the population's attitude or opinion with time is one of *kind*, rather than *degree*. People have not simply become "more" or "less" favorable; they have changed their definition of the issue. To measure a change in *degree* of "favorableness," it would appear to be necessary that the attitude remain in the same continuum both times. This hypothesis, if correct, indicates a basic weakness in much of the present trend analysis based upon repeating a single question. Unless it is known that the meaning of this question has not changed with time, the statement that a group has become "more" or "less" favorable is a doubtful interpretation. Scale analysis affords a way to test whether or not the content of the question has retained its original meaning.

An especially important example of this problem in trend analysis applies to experimental studies of induced change, i.e., the effectiveness of propaganda by means of before and after group tests. Unless the same scale is obtained before and after exposure, one should speak of changes in "degree" of favorableness with great caution. This is especially evident when, for example, an item such as "All Russians are Communists," which *before* the exposure scaled with other attitude-toward-Russia items, shows a great change but no longer scales. One hypothesis could be that this item has changed from an attitude universe to an information universe, due perhaps to the stress of the propaganda exposure on the facts of the case. A test of this hypothesis might be to include other informational items, before and after, and see if those items which appear to shift from attitudinal to informational, no longer scale with the attitudinal items but do scale with the informational items.

* * * * *

The ramifications of scale analysis discussed above constitute a bare minimum of all the problems of scale analysis which remain to be fully investigated. The present research represents only an initial attempt to formulate the foundations for the present theory of scale analysis and to develop a workable technique for the determination of scales by means of this theory. As has been stated previously, the present research is based upon data gathered in the course of the war and is limited almost entirely to the attitudes of soldiers.

The further testing of the present theory with other types of social data is an important avenue for future research. It can be expected that future research will serve to show more clearly than is presently possible both the disadvantages and advantages of the present theory and method.

CHAPTER 6

RELATION OF SCALOGRAM ANALYSIS TO OTHER TECHNIQUES[1]

VARIOUS approaches to the study of attitudes and opinions have been made in the past. These approaches differ occasionally in objectives. In order to obtain a further perspective on the purpose and techniques of scalogram analysis, it may be helpful to compare this approach with previous ones. Emphasis will be placed primarily on similarities or differences in purpose and usefulness. We begin with a discussion of the distinction between scale and nonscale approaches to the study of problems. We then compare scale analysis with several other means of analyzing universes of data, e.g., the use of arbitrary indexes, item analysis, equal-appearing intervals, paired comparisons and factor analysis.

Scale and Nonscale Approaches

In order to understand better what types of problems are best studied by means of scales, it may be helpful to have a clear idea as to what types are probably not profitable for scale analysis. In the analysis of qualitative data, a vast number of problems definitely do not involve the notion of a scale. A great many nonattitudinal problems involving qualitative variables, like sex, religion, etc., usually deal with each attribute as a complete variable in itself, and not as representing a universe of similar items. Such cases are *not* scale problems. On the other hand, many problems in social psychology regard a qualitative item, like an opinion or attitude statement, as but a sample from a universe of similar items. Here is where a scale analysis is helpful. It tells whether or not the universe is scalable by using only a small sample of the items. Social phenomena are usually complex. However, if a scale is found to exist for a universe of phenomena, that means that a certain simplicity attends those phenomena. The theory of scales tells how

[1] By Louis Guttman.

to recognize and take advantage of that simplicity. If the phenomena do not follow a simple pattern, then scale analysis shows that a more complicated technique is needed to handle the data properly.

A problem involving a universe of data, therefore, should first be subjected to a scale analysis. The analysis will show either that scales are involved or are not involved. A problem may be thought a priori to involve a scalable universe, but analysis of the data may show that a scale does not exist. The existence or nonexistence of a scale is not a criterion of the worth of a problem. If a problem turns out to involve nonscalable data, they should be treated, not as scales, but in whatever manner will yield a proper answer.

To summarize briefly, problems which do not involve samples from a universe of items are not in general scale problems. Problems which do involve such sampling—including almost all of attitude and public opinion work—will profitably be studied first by means of a scale analysis. If the universe is scalable, or can be broken down into scalable subuniverses, then it can be handled very easily by simple scale scores. How best to handle nonscalable universes remains a far more complicated, and as yet unsolved, problem.

The distinction between scale and nonscale data may perhaps become even more clear when we consider the different roles that a factor or variable can play in a given research project. Research usually has either or both of two purposes: description and prediction. Let us start with prediction.

Two roles for variables in prediction studies. Each variable in a prediction study plays one of two possible roles. It can serve either as something to be predicted—in which case it is often called technically a *criterion*—or it can serve as a predictor. There are many possible ways in which each of these two roles can be assumed, such as in multiple and in partial correlation, but this should not be allowed to obscure the fundamental nature of the roles.

Each of these two roles can be filled by qualitative variables (attributes) or by quantitative variables. Conventional statistical textbooks concern themselves largely with the prediction of quantitative variables. If the predictors are also quantitative, then linear or curvilinear regression equations are derived for testing predictability. If the predictors are qualitative, then the approach of analysis of variance is used.

Relatively little attention has been given in the past to the case

where qualitative variables are the thing to be predicted; and very little attention has been given to the notion of a *universe* of qualitative variables, either as a criterion or as a predictor. Qualitative criteria need techniques different from those of quantitative variables. Least squares and analysis of variance do not apply. Furthermore, the study of *universes* of variables requires a special approach. Scale analysis studies a universe of qualitative data in a method appropriate to such data. It proves that, if the universe is scalable, it can be handled very easily in either role, whether as criterion or as predictor, by a simple scoring scheme. On the other hand, if the universe is not a scale (nor a quasi scale), then use of an arbitrary scoring scheme will not in general permit the universe to fill either role properly.

The problem for analysis, then, is to ascertain the complete multivariate distribution of an indefinitely large universe of attributes. This means finding out how each respondent stands on each of the possible items. Here is where scale analysis comes in. If the universe of items is a scale, then each respondent's pattern of responses can be represented by his scale rank among all respondents. From the point of view of description, an important feature of a scale is that a man with a higher scale score than another will be just as high or higher on every item in the universe (within scale error). He is known (within scale error) to be in the one answer category of each question that covers his scale rank. From the point of view of prediction as well as description, a crucial feature is that the distribution of scale ranks *based on only a sample of items* is known to represent the complete multivariate distribution of the universe of items. A relatively few items can be used from a scale to represent the universe accurately either as a criterion or as a predictor.

The Use of Arbitrary Indexes

Although techniques for scale analysis are very simple, the question may be raised as to whether or not they can be dispensed with entirely. If a set of items is agreed to have a homogeneous content, to belong to the same universe, of what advantage is scale analysis? Why bother with determining the reproducibility of the area? Why not simply assign arbitrary weights to the responses to each question, add up these weights to obtain a total score which can be called an *index* of the area in question? Cannot such an index be adequate for relating this attitude area to other behavior? This is often the solution reached by public opinion analysts who ask a "battery" or

series of opinion questions on what appears to be the same issue and then proceed to count up the number of "favorable" replies.[2]

It is true that if an area is scalable, then the resulting scale scores will correlate very highly with any index obtained by arbitrary weights, provided the weights are in the right direction. Scale theory proves that there is no harm in obtaining an apparently arbitrary index from a scalable area, either for descriptive or predictive purposes. That is why only very simple weights and scores are used in scale analysis in practice. But if the area is *not* a scale (or quasi scale), then we have an entirely different story.

If an area is not a scale, then it may often be possible to break it down into subareas of content that are scalable separately. In such a case, each person should receive *several* scale scores, one on each subarea. The entire area can then be properly described by use of these several sets of scores. These scores can serve as the criteria if it is desired to predict the area from other variables; it is the *profile* of scale scores that is to be predicted in such a case. Similarly, if it is desired to use the items in the area as predictors of an outside variable, it is necessary only to determine the multiple correlation of that outside variable on the scale scores (which is very little work compared to computing directly the multiple correlation on all the items themselves).

Omitting a scale analysis and just going ahead with a single arbitrary index can completely obfuscate the purpose of the research, whether for descriptive or predictive purposes, if in reality several scores are required and not just one. There are at least five basic defects to using an arbitrary index if the area is not actually a scale (or quasi scale). These are: (1) Lack of descriptive meaning. (2) No criterion for weights. (3) Sampling of items and description. (4) Improper weights for prediction. (5) Sampling of items and prediction. Each of these will now be discussed in turn.

1. *Lack of descriptive meaning.* A first defect concerns the descriptive meaning of the index. In a scale, each score has a definite meaning with respect to the actual items. From a person's score can be reproduced his response to each and every question (within scale error). As a consequence, it is meaningful to say that a person with a higher score than another has a more favorable attitude than the other person; this is true because a higher score means an

[2] See, for example, F. Mosteller and H. Cantril, "The Use and Value of a Battery of Questions," in H. Cantril (editor), *Gauging Public Opinion* (Princeton University Press, Princeton, N.J., 1944), pp. 67–73.

equivalent or higher response to each and every item (within scale error). This simple and direct meaning for scores does not at all obtain for nonscalable areas. No single index can reproduce a person's responses to the items in such a case, so that no single index can be said to be an adequate representation of the area. Two persons can have the same score from quite different behavior. Furthermore, the fact that one person has a higher index score than another person has no particular meaning with respect to the content involved; the first person may be more favorable on some items but less favorable on other items than the second person. There is no clear meaning as to what a "higher" attitude is, nor even as to what "equality" on the attitude signifies.

For example, one person might say, "I think Negroes should have a right to vote, but I don't think they should be given skilled jobs." Another might say, "I don't think Negroes should have a right to vote, but I do think they should be given skilled jobs." Which person has the more favorable expression of attitude toward Negroes? Finding these two kinds of people implies that the area is not scalable. As a consequence, no index score based on arbitrary weights will provide a meaningful ranking of attitudes of the population. As more and more items are added to these two on voting and jobs, and if the additional items continue to depart from a scale pattern so that many nonscale types of persons are found, any index score will have less and less content meaning; it will be less and less able to reproduce the responses to the items.[3]

2. *No criterion for weights.* The lack of meaning of "equality" or "more favorable" and "less favorable" attitudes becomes even more evident when it is recalled that the weights in obtaining an index are arbitrary, which leads to a second basic descriptive defect of such an index for a nonscalable area. By changing the weights, persons who are "equal" according to one set will no longer be "equal" according to the other. Persons who were "higher" than others according to one set of weights may turn out to be "lower"

[3] A good example of the distinction between arbitrary scoring and scale analysis comes from some unpublished data of Paul Wallin. He was studying adjustment in engagement. The items in the questionnaire were all judged to be homogeneous in general content, each item expressing some relevant aspect of adjustment in engagement. Yet scale analysis showed that the items should be divided into two subareas of content, one of which was a scale and the other a quasi scale. It was most striking to find that the scores on these two subareas *correlated zero* with each other! This proves that it is rather meaningless to speak of general adjustment in engagement as a single variable, which an arbitrary scoring of the original questionnaire would have implied; an arbitrary index based on the combined areas would represent neither area.

on the other set. Which set of weights are the ones to be used? Shall the "right to vote" obtain a higher or lower weight than the "right to a skilled job" in attitude toward Negroes? There is an infinite variety of weights which can be chosen arbitrarily for an index, each of which may seem "reasonable." If the area is not a scale, index scores obtained from different weights can have relatively low correlations among themselves. This is especially true where two subareas correlate zero with each other. What should the research worker do when confronted with several sets of weights in such a case? Without a frame of reference, such as provided by scale analysis, there can be no solution to the problem of what is best to do.

3. *Sampling of items and description.* A third defect to omitting a scale analysis before obtaining descriptive ranks or scores for people arises from the problem of selection of items. Just as index scores can change in nonscalable areas when different arbitrary weights are used, in the same way they can change if different items from a nonscalable area are used. If fifty items are selected from the area, then an index score on the first twenty-five can correlate very slightly with an index score on the second twenty-five if the area is not a scale, and the index score on the whole fifty can differ widely from the index scores that would have been obtained if a different fifty questions had been used. This is especially true in the case of public opinion questions which consist usually of a series of less than ten questions. The correlation between indexes can be zero. There is no rationale for defining what is meant by ideal scores on the indefinitely large universe of items from which the sample of items were taken in such a case.

One of the important features of scales is that this problem of sampling of items is minimized. If an area is a scale, then any sample of items from the area will yield essentially the same rank order for the people as any other set of items. Respondents to a public opinion poll will be ranked the same regardless of the specific questions asked. There is an ideal rank order defined for the infinite universe of items, and any sample of items from that universe will yield ranks that closely approximate the universe ideal. Knowing that an area is a scale means that there need be no great concern about the sampling of items; but knowing the area is *not* a scale means that there must be great concern about the sampling of items. Remaining in ignorance as to the scalability of an area means remaining in ignorance of whether or not to be concerned about the

sampling of items, and leaves the description open to serious criticism.

4. *Improper weights for prediction.* The problem of sampling of items and the problem of weights are both felt with full force simultaneously when it comes to problems of *prediction.* A fourth important defect of the suggestion that a scale analysis is unnecessary concerns the use of the items in the role of predictors. If it is desired to predict an outside variable from a given set of items, then the best technique for doing this takes into account all the joint relationships within the items and between the items and the thing to be predicted. The best possible prediction, based on this complete information, is called multiple correlation or the discriminant function technique, depending on whether the thing to be predicted is quantitative or qualitative.[4] Let us restrict our discussion here to the linear case where the best prediction can be obtained by weighting the items and obtaining simple sums as prediction scores. If the set of items does not belong to a scale, then *a new set of weights must in general be used to predict each different outside variable.* The items must be weighted one way to predict one criterion and another way to predict another criterion if the best prediction in each case is to be obtained. The use of arbitrary weights to obtain an index score necessarily means, then, that the predictive power of items is not being realized. An arbitrary index in general will *underestimate* the predictability of any criterion from the items. The index can correlate practically zero with a given criterion, whereas, if the actual multiple correlation or discriminant function were worked out, the items could be found to correlate very highly with the criterion.

Suppose a set of items on the role the United States should play in international relations proved *not* to be scalable. Suppose further that arbitrary weights were assigned to the items anyhow, and the respondents were scored according to the resulting arbitrary index. And suppose further that this index was found to have a correlation coefficient of .3 with, say, attitude toward England, where the latter attitude was scalable. It would be quite fallacious to say that attitude toward England is predictable from attitude toward international relations only to the extent permitted by a

[4] For a brief outline of multiple correlation and the discriminant function, see Louis Guttman, "An Outline of the Statistical Theory of Prediction," in P. Horst, et al., *The Prediction of Personal Adjustment* (Social Science Research Council, New York, 1941), pp. 253–312.

correlation of .3. The arbitrary weights used for the index are almost certainly not the best for the purpose of predicting attitude toward England. If the best weights for this particular prediction purpose were computed by the technique of multiple correlation, the coefficient of multiple correlation could be much higher than .3. Instead of being negligibly predictable, attitude toward England might be highly predictable from the items in the area on international relations. The arbitrary index scores can be completely misleading concerning the extent of this predictability.

In the case of a scalable area, the multiple correlation or discriminant function must always be given very approximately by the scale scores, *regardless of what is being predicted.* If a correlation with the scale scores is .3, then the multiple correlation coefficient with the separate items is necessarily very close to .3. Whether predictability is high or low will, of course, depend on what is being predicted from the items; but regardless of what is being predicted, maximum predictability from the items is given by the scale scores. Knowing that the set of items is scalable means that it is known that the full predictive power of the items is being used in any situation. Knowing that the items are not from a scale means that it is known that the items must in general be reweighted for each new prediction problem. Not knowing whether or not the area is a scale, but just using arbitrary weights to obtain index scores means knowing that in many cases much of the predictive power of the items is being thrown away by the index.

5. *Sampling of items and prediction.* A fifth drawback to not finding out the scalability or reproducibility of an area concerns the relationship of multiple correlation to the sampling of items. If the area is scalable, then it is known that not only the multiple correlation of the thing being predicted with the set of items is essentially the correlation of that thing with the scale scores, but also it is known that *adding more items* to the sample from the same scalable universe will *not* increase the multiple correlation. That is why relatively few items can be used from a scalable area for prediction purposes, yet the predictive power of the infinite number of items from which this sample was drawn is being fully realized. For example, only five or six items from each of fifteen areas were used in a study of psychoneurotics in the Army. Using five or six hundred items in each of these areas instead of just the five or six would not have increased the multiple correlation.

This feature of scale analysis is of utmost importance in public opinion polling, where the problem of prediction from as small a number of questions as possible is of major practical importance.

On the other hand, if an area is not a scale, then adding more items will in general *increase* the multiple correlation. There is no way of knowing when the maximum correlation has been reached. The possibility remains that more items in the same area can still help in a given prediction problem. There is as yet no way of determining what the multiple correlation would be on the infinite universe of items from only a sample of items if the area is not a scale (or a quasi scale). That means that one cannot speak of the predictability of an outside variable from the items except as an understatement. The multiple correlation on the universe of items is certainly as big as that obtained in the sample but may be much bigger. A sample of items may yield just a small multiple correlation whereas in fact the universe may yield a very high correlation.

Knowing that an area is scalable means that the problem of sampling of items for prediction purposes can be simply handled; the maximum predictability in the area can be determined just by means of a sample of items. For nonscalable universes, there is the ever-present possibility that the predictive power of the whole area is far greater than the predictive power of the sample of items used; and multiple correlation based on a sample of items from a nonscalable universe can be very misleading. It follows that a correlation based on an arbitrary index derived from the sample will be even more misleading, because this correlation will underestimate in turn even the sample multiple correlation.

The successive shrinkage in predictability due to the sampling of items and to the use of an arbitrary index for nonscalable areas can be visualized as follows:

Multiple correlation on a nonscalable universe of items	>	Multiple correlation on a sample of items from the nonscalable universe	>	Correlation with arbitrary index of sample of items

The retention of the full predictive power of a scalable universe by a sample of questions, on the other hand, can be visualized as follows:

Multiple correlation on a scalable universe of items	=	Multiple correlation on a sample of items from the scalable universe	=	Correlation with sample scale scores

Item Analysis

Similarity of purpose. The purpose of scale analysis is to test the hypothesis that a universe of items comprises but a single factor in the sense that from but a single set of scores (on this factor) the responses to each of the items can be reproduced. The widely used technique of item analysis often has a similar descriptive motivation in that it is desired to test whether or not a set of items is related to but a single factor. However, important distinctions exist between the scaling and the item analysis approaches in at least two basic respects: one is the manner in which the objective is defined, and the other is the resulting difference in techniques.

If both approaches are regarded as being directed toward the same problem of describing internal consistency, then it will be found in the following discussion that at times item analysis may entirely miss the mark. This may seem like a radical criticism of item analysis in view of the fact that the technique is so widely used in the literature. The criticism will not seem so radical,[5] however, when item analysis is viewed in its historical perspective. Item analysis, with respect to attitudes, seems to have originated in a different context from that of exploring the structure of a set of items. It was in the field of predicting an outside variable that item analysis was initiated and it is in this field that it is actually helpful and proper to a large extent. When the newer field of attitude analysis was developed, item analysis was borrowed rather uncritically in order to describe the internal consistency of attitude items.

The present theory provides a coherent approach to scale analysis from the point of view of internal consistency. We will attempt to show that the borrowing of the item analysis technique is no longer advisable.

Proper use for a single prediction problem. Item analysis is often useful in a situation where it is desired to use the items as predictors of a given outside variable. The best way to accomplish this pre-

[5] This criticism is actually not original here. It was made many years ago in the mental testing field, where much more attention has been given to the rationale of analyses than in the attitude testing field.

diction is by means of multiple correlation or the discriminant function. But to do the complete multivariate analysis required by these techniques is a laborious, time-consuming, and often prohibitive undertaking when many items are involved. Item analysis has been shown empirically to be able to provide a short cut in selecting a relatively small subset of items that will predict the criterion with substantial accuracy even though it is known that predictive efficiency from the whole composite is lost thereby. Item analysis is carried out in practice by various techniques; but they all consist of relating each item individually to the external variable, and then selecting and weighting each item according to its separate correlation. The prediction variable arrived at by means of item analysis will not be as good in general as the one arrived at by the complete multivariate analysis required by multiple correlation or the discriminant function, but it may sometimes be adequate for the investigator's purpose.[6] It must be remembered, though, that if the same items are to be used to predict some other criterion in a new problem, then a completely new item analysis must be performed on all the original items. A new selection and weighting must be made in general for each new prediction problem.

Improper use for studying internal consistency. The use of item analysis for saving work in a prediction study can be quite proper and helpful. It is quite another matter, however, to use item analysis as a technique for studying the internal consistency of a set of items. The problem of describing internal consistency can eventually tie in with the problem of prediction, but it must do so without reference to any particular outside criterion. A scalable area, for example, *because* of its pattern of consistency, can be used in a very simple fashion to predict *any* outside variable whatsoever; the simple scale scores are invariant quantifications of the items for predicting any external criterion. But the problem of describing internal consistency itself must not be confused with the problem of external predictions. According to scale theory, the problem of internal consistency is stated to be: Can each item inside the area be reproduced from scale scores? The amount of reproducibility is computed directly by a simple counting up of errors; this is a proper

[6] Throughout this paragraph, it is assumed that a large sample of persons is used for the computations. If a relatively small sample of persons is used, item analysis may be *better* than multiple correlation when it comes to predictions for a new sample of persons. This has been shown theoretically in Louis Guttman, *ibid.*, and it has been shown empirically in Louis Guttman, "Two Studies in Weighting Techniques," *op. cit.*, pp. 349–364.

way to measure the correlation of a qualitative variable with any other variable.

Item analysis, as employed now for the study of internal consistency of a set of items, does not seem to have such a clear rationale in mind. The reasoning seems to be, instead, somewhat as follows. The items are all thought to be indicators or measurements of some central variable. The problem is to obtain persons' scores on this variable. If this central variable could be observed by some external means, then of course a good way of estimating persons' scores through the items would be to work out the multiple correlation of this variable on the items; and in order to save the labor of multiple correlation, item analysis might be employed to select and score the items to obtain total estimates of the central variable. If there really were such a variable which could be defined apart from the items, this statement of the problem resolves again into prediction of an outside variable. But unfortunately, research workers do not observe a central variable external to the items; and now reasoning seems to be of the following nature. Since there is no external variable with which to perform an item analysis, why not obtain a variable from the items themselves? This can then be used in lieu of an external variable, and an item analysis can then be performed.

To obtain a variable from the items themselves as a starting point, arbitrary weights are assigned to the items, and a total score is obtained by summing the weights for each person. These total scores are used as though they were external to the items, and then the relationship of the total score to each item is examined by any of several item analysis techniques.

It would, of course, be foolish to work out the complete multiple correlation of the total score on the items, since it is known in advance that it must be perfect; the total score is a perfect function of the items, being an exact sum of them. What sense, then, does item analysis have in this context of internal consistency? It cannot be regarded as an approximation to multiple correlation because there is no need for such an approximation in this case where the multiple correlation is known (to be perfect). Does the item analysis accurately describe the internal consistency of the data then? The answer again must be in the negative, if by internal consistency is meant the reproducibility of the items from the total score.[7]

[7] This has also been recognized earlier in the field of mental testing. In criticizing the use of various critical ratios (which are typical techniques in attitude item analysis)

Item analysis does not attempt to see how well items can be reproduced from the total score, but rather it attempts to do *just the opposite;* item analysis investigates how well the total score can be estimated from each item separately. This is a crucial distinction. Scale analysis regards the score as a representation of the items; the score is to be a means of representing the items in any situation, which will be possible if each of the items is perfectly related to the score. The *converse* is done by item analysis; here, the item is supposed to represent the score.

Being reproduced versus discriminating. It can easily be seen that even though the items form a scale, so that each item is a perfect function of the score, nevertheless the score can be negligibly correlated with any one item. For example, consider a dichotomous question from a scalable universe in which 60 per cent of the people said "Yes" and 40 per cent said "No," and where "Yes" is more favorable than "No." Then it must be that everybody in the top 60 per cent of rank order on the entire scale said "Yes" to the question and everybody in the bottom 40 per cent said "No." Each person at a given rank has the same response to the question. The converse is not true. To each response to the question correspond many ranks. If a person said "Yes" to the question, we *cannot* know his rank very closely from this information alone; it can be anything between the fortieth and one hundredth percentile. If he said "No" his rank could be anything between the zero and fortieth percentile. This is not a very high correlation of the rank with the response, whereas the correlation of the response with the rank is perfect. Item analysis would discover that the mean rank of those who said "Yes" is different from the mean rank of those who said "No" (or some equivalent relationship if ranks are not used). *Finding such a mean difference says little about the reproducibility of the item from the rank.* A large mean difference between ranks can be found for nonscalable items, as well as for scalable items. Testing to see whether an item will *discriminate* between total scores has little to do with testing to see whether an item can be *reproduced* from a total score. Item analysis will find that items "discriminate" regardless of scalability. All that is required is sig-

in intelligence item analysis, Guilford states that "the amount of overlapping of high and low groups . . . gives a truer picture of discriminatory power." J. P. Guilford, *Psychometric Methods* (McGraw-Hill Book Co., Inc., New York, 1936), p. 434. He uses the word "validity" in this context where we would use the word "scalability." Studying the overlapping of high and low groups is tantamount to studying reproducibility

nificant correlation ratios of scores on items, which says nothing at all about reproduction of items from scores.

It seems clear, then, that if it is intended to study the internal structure of a set of items, item analysis is quite beside the point. Item analysis should be used only in a situation involving an external criterion, where the items are intended to be only aids in this one prediction problem. It is misplaced when used in connection with internal analyses.

Equal-Appearing Intervals and "Cafeteria" Questions

The procedure of equal-appearing intervals. One of the most widely used approaches at present for the descriptive analysis of attitude items is the method of equal-appearing intervals as developed largely by Thurstone.[8] This approach differs considerably in its rationale from that of scale analysis. Indeed, the two theories can be considered as each being devoted to a distinct and separate problem, and each is appropriate for its own purpose.

Briefly, the procedure for equal-appearing intervals is to have a group of judges rank a set of statements in the way they think will best fit the population of individuals to be studied. The items are always in a dichotomous form. A declarative statement like "The British are fine people" is made, and the response of each subject to whom the questionnaire will eventually be administered is to be either an endorsement or lack of endorsement of such a proposition. The task assigned the judges is to determine weights for the endorsements of a series of such statements. This is usually done by having the judges rank the statements into eleven piles. If a judge believes that endorsing a statement means a very favorable response, then he will place it in a pile with a high rank. If he believes an endorsement means a very unfavorable response, he will place it in a pile with a low rank.

This ranking into piles is done by a group of judges. The consistency of the ranking of the judges is then analyzed from two points of view. One is to reject items which are deemed to be ambiguous or otherwise faulty because of wide disagreements in the rankings of the judges. The other is to assign weights to the statements. The weight given to a particular statement is usually the median of the rankings assigned to it by the judges.

The judges assign the weights, not according to how they them-

[8] See L. L. Thurstone and E. J. Chave, *The Measurement of Attitude* (University of Chicago Press, Chicago, 1929).

selves would respond to the questions, but rather in terms of how strong an opinion or attitude they believe endorsement of the statements would imply for the subjects who are to take the questionnaire. The selection of items and weights, therefore, is intended to reflect the judgments of the judges as to the role that endorsement of each proposition plays in the attitude of the subjects.

The final selection of items is then administered to a sample from a population of individuals. The statements that each individual endorses have their weights added up, and the average represents the individual's score on the attitude represented by the items. A person with a higher score is said to have a higher attitude, and a person with a lower score is said to have a lower attitude.

Difference in rationale. In so far as the consistency of judgments of the judges is being studied, the problem of equal-appearing intervals is quite different from that of scale analysis. In scale analysis, there is no concern with the behavior of judges. The entire concern is with the behavior of the ultimate respondents. It is the consistency of the respondents that is studied by scale analysis and not the consistency of judges' judgments. It has been found in at least several studies that there is no necessary relationship between the weights assigned by judges and the pattern of behavior of the respondents.[9] From a person's score obtained by the method of equal-appearing intervals, one cannot in general reproduce his response to each of the questions. Two people with the same score can have quite different patterns of responses. In other words, selection and weighting of items according to equal-appearing intervals does not at all necessarily yield a scale in the sense of reproducibility from scale scores.

Difference in utility. As a corollary to the fact that treating items by the method of equal-appearing intervals does not necessarily yield a scale of respondents' behavior, all the problems concerning nonscalable areas also can attend these items. If it is desired to predict an outside variable from the items, then using the weights given by judges will in general underestimate the correlation that would have been obtained if the specific multiple correlation weights were used instead. The scores obtained from judges' weights, in so far as they reflect the behavior of the judges and not of the respondents, do not represent the responses of the respondents and

---

[9] See, for example, references in Hugh Carter, "Recent American Studies in Attitudes Toward War: A Summary and Evaluation," *American Sociological Review*, Vol. 10, No. 3 (June 1945), pp. 343–352.

hence are not an invariant quantification of the respondents' answers for external purposes. Predictive power of the items will in general be lost by the use of judges' weights.

The method of equal-appearing intervals is helpful in studying the judgments of people in comparing various things. It can be regarded as a variation on paired comparisons (which is the next technique to be discussed below). It does not seem very appropriate for studying the internal consistency of the actual responses of respondents. Consistency of judges and consistency of respondents are two distinct and separate problems.

"Cafeteria" questions. The lack of correspondence between judgments of responses and scalability of responses can also be found in a technique related to that of equal-appearing intervals. A common device used in opinion polls is the "cafeteria" type of question. Here, several statements are prepared which are often judged to represent different degrees of favorableness on an issue, and the respondent is asked to indicate which statement comes closest to expressing his opinion. For example, one such question used recently in national public opinion polls is the following:

> With which *one* of these statements concerning postwar relations with Russia do you come closest to agreeing?
>
> 1. It is very important to keep on friendly terms with Russia, and we should make every possible effort to do so
> 2. It is important for the U.S. to be on friendly terms with Russia, but not so important that we should make too many concessions to her
> 3. If Russia wants to keep on friendly terms with us, we shouldn't discourage her, but there is no reason why we should make any special effort to be friendly
> 4. We shall be better off if we have just as little as possible to do with Russia

The polling agencies using this question apparently judged that endorsing the first statement implied a more favorable attitude than endorsing the second, etc. The "cafeteria" question differs from the more complete treatment by equal-appearing intervals in that only a rank order of categories is attempted, and not a more precise metric.[10]

In order to test the hypothesis that the four statements actually indicate degrees of favorableness, the Research Branch asked the

[10] In a study reported by Cantril, judges were used to rate the various statements in a "cafeteria" question with the purpose of securing measurable distances between the different alternatives. See Hadley Cantril, "The Intensity of an Attitude," *Journal of Abnormal and Social Psychology*, Vol. 41, No. 2 (April 1946), pp. 129–135.

question on Russia of a sample of 3,000 enlisted men in the following form:

> Do you agree or disagree with this statement?
> It is very important to keep on friendly terms with Russia, and we should make every possible effort to do so.
> _____ Agree
> _____ Disagree
> _____ Undecided

Similarly, each man was asked if he agreed or disagreed with each of the remaining three statements in the original "cafeteria" question. Thus the original question was broken up into four questions which could then be tested for scalability. A scalogram analysis showed that these four questions did not form a scale. This would indicate that more than one dimension was involved in the cafeteria answer alternatives and consequently the question could not be used to rank respondents according to degree of favorableness toward Russia.

A further comparison was made by asking the "cafeteria" question as originally worded following the rewording into four "agree-disagree" questions. Of all men who checked statement 4 (the most extremely anti-Russian statement) when asked in the cafeteria form, 49 per cent also "agreed" with statement 1 (the most extremely pro-Russian statement) when presented as a separate question. Again, of all men who checked statement 4 in the cafeteria form, more (77 per cent) "agree" with statement 2 than with statement 3 (59 per cent) when presented as separate questions, which would indicate that statement 3 is less extreme than statement 2. These results confirm what was proved by the scale analysis, namely, that more than one dimension of response is being tapped by the cafeteria question.

Judgment of the dimensionality of a "cafeteria" question with respect to the actual responses seems to have the same basic weakness as judgments in the more complete equal-appearing intervals technique. A study of the interrelationships of the responses themselves seems essential to determine whether or not a meaningful rank order is present.

Since a "cafeteria" question obtains but one response from each person, it permits for no check on dimensionality. It seems advisable in general, then, to ask a series of "agree-disagree" questions (or some similar form), rather than to combine them into a single question. This provides several answers for each respondent, so

that a scale analysis can then be performed to test the hypothesis that but a single dimension is present.

Paired Comparisons

Difference in purpose. The purpose of scalogram analysis is to find out whether or not the attitude of people toward *one* thing (one universe of items) can be represented by a single rank ordering of people. Paired comparisons has a related but different purpose; it attempts to determine average differences between the population's attitudes toward *several* things. Scale analysis tests the hypothesis that it is meaningful to say that one person has a higher attitude than another toward some one object. Paired comparisons tries to find out if the average attitude of a population toward one object is higher than its average attitude toward another object. Hence the goals of scale analysis and of paired comparisons are quite distinct.

An important use of paired comparisons made by the Research Branch, for example, was with respect to the point system for demobilization.[11] The problem was to assign weights to the four factors considered important by the soldiers with respect to determining who should be released from the Army first: length of time in the Army, length of time overseas, amount of combat, and number of children. Should a battle receive less weight or more weight than a baby? Should a month of service overseas count the same as a month of service in continental United States? These were the kinds of questions to be resolved in the light of the attitudes of the enlisted men.

No single set of weights for these categories would be able to reproduce each man's judgments, because people would be found who would give opposing judgments. Men in continental United States would tend to discount overseas service and combat; men overseas would tend to discount anything involving United States duty. Family men would tend to emphasize children; nonfamily combat men would tend to emphasize battle credit and discount children.

If we wished, we could make a separate study of each of the four factors. A series of questions could be asked concerning, say, children alone. A scale analysis would test the hypothesis that attitude toward children, with respect to demobilization, was a scale. Similarly, attitude toward each of the remaining three factors could

[11] See Louis Guttman, "An Approach to Quantifying Paired Comparisons and Rank Order," *Annals of Mathematical Statistics,* Vol. 17 (1946), pp. 144–163.

be tested separately for scalability. Studying attitudes toward each of the separate factors would be a scale problem, so that four distinct scale analyses could have been made.

Paired comparisons, however, is not directly concerned with how people are ranked on each of the separate factors. It is not concerned with saying that one person has a more favorable attitude than another with respect to credit for children. Instead, it is interested in an *average comparison* of the attitudes on the factors; e.g., on the average, are men more favorable toward credit for children than toward credit for battle?

No single factor, or no single system of weights, will account for all the judgments of the men on the four demobilization factors. The problem of paired comparisons is usually a multifactor problem with respect to individual responses. People offer different systems of judgments according to their different attitudes toward each of the things being judged.

If each of the four attitudes in the score card problem were scalable, and if these four scales were not perfectly intercorrelated (which undoubtedly would be the case for these data), then it would be necessary to give each man *four* scores in order to reproduce his judgments, one score on each attitude. The final weights assigned by paired comparisons to an object would represent an *average* of the scores of the population on that object. If one object receives a higher weight than another, that means that on the average the population's attitude toward the first object was higher than on the second.

Paired comparisons inquires directly into the average differences between attitudes toward the things being compared, without any analysis of the internal structure of the separate attitudes (which is the problem of scale analysis). An average difference, like any other average, must be considered in the light of the dispersion about the average. The less correlated the attitudes are, the more dispersion there will be in the comparisons, and the less reproducible are the judgments of each person from knowledge of only the group averages. The system of weights finally adopted for the score card is approximately that which would best help reproduce the set of judgments of each of the individual soldiers. It will contradict many of the judgments, but on the whole it is the best fitting single system of weights. Any different set of weights will contradict even more judgments than the weights used.

From paired comparisons, then, one attempts to reproduce the

judgments of the people from the weights of the objects. In contradistinction, scale analysis assigns each person a rank and from that reproduces his responses.

A perfect scale would have one and only one factor, so that reproducibility would be perfect. Paired comparisons is usually not a single factor problem; all people rarely rank the objects in the same way, because their attitudes toward the things being compared are generally not perfectly intercorrelated. At best, paired comparisons can reproduce the *proportion* of people who say one object should be higher than the other. The method of paired comparisons, as developed by Thurstone and others,[12] focuses entirely on reproducing proportions of judgments. It introduces an assumption of a normal distribution for the attitudes, and then proceeds to see if, by using this assumption, weights can be devised for the objects being judged which will reproduce the table of proportions of judgments of each kind. The newer approach to paired comparisons, as used in the score card problem, does not make any assumptions about normal distributions or any other kind of distributions, but focuses rather on reproducing each individual's judgment, not just the group proportions of judgments. In either case, it is the objects whose ranking is analyzed (in terms of average attitudes), and not the people. In scale analysis, it is the people who are ranked and not the objects. In paired comparisons, the weights for objects are averages which usually cannot reproduce individual judgments very closely. In a scale, from a person's rank one can reproduce his responses to each item almost perfectly.

Factor Analysis

Factor analysis designed only for quantitative variables. In the field of mental testing, there has been developed over a number of years an approach to analyzing the structure of a system of many quantitative variables. Originated as a single factor theory by Spearman, and developed into a multifactor theory by Thurstone and others, this approach is now being widely used in analyzing intercorrelations of many kinds of sets of variables apart from mental tests. If we are concerned with a universe of qualitative items, i.e., a large set of qualitative variables, could not factor analysis serve as well to analyze the structure of this universe? What relationship would there be between results obtained from a Spear-

[12] For a discussion of the work of psychologists on paired comparisons see J. P. Guilford, *op. cit.*, especially Chapter 7.

man-Thurstone factor analysis of a set of qualitative items and results obtained by a scale analysis of those items?

Briefly, the answer to this question can be stated as follows: a factor analysis in the Spearman-Thurstone sense will fail to test adequately the scalability of qualitative data, because it was not designed for that problem. Scale analysis is designed as a single-factor theory for qualitative data. *From a scale analysis it can be known what a factor analysis will show.* The converse is not true; from a factor analysis it will usually be difficult, if not impossible, to know what a scale analysis will show. Scale analysis, as outlined previously, makes a complete analysis of qualitative data, using no extraneous assumptions, and using only techniques appropriate to qualitative data. A factor analysis in the sense of Spearman and Thurstone was not designed for qualitative data and will not make a complete analysis of qualitative data; it will ordinarily use extraneous and often misleading assumptions, will not use techniques appropriate for qualitative data, and can lead to quite erroneous interpretations.

The need for distinguishing qualitative from quantitative variables. Factor analysis in the Spearman-Thurstone sense was devised for the study of *quantitative* variables. When variables are quantitative they can often have linear regressions, in fact can have normal distributions, and can have their interrelations measured by product-moment correlation coefficients. It is well known that in a normal multivariate distribution, all multiple and partial correlations between any subsets of variables can be completely expressed in terms of zero-order correlations. Therefore, in studying quantitative variables, especially those which have a normal distribution, it is sufficient usually to consider only the correlations between the variables two at a time. That is why, in factor analysis, only the matrix of zero-order correlations is analyzed. Higher order relationships are implicitly analyzed if the zero-order structure is known. If r_{xy}, r_{xz}, and r_{yz} are known, then immediately the following correlations are also known: $r_{xy.z}$, $r_{xz.y}$, $r_{yz.x}$, $r_{x.yz}$, $r_{y.xz}$, and $r_{z\,xy}$. Similarly, for the case of more than three variables, all partials and multiples are known from just the zero-order correlation coefficients.

When it comes to qualitative data, two of the basic features are lacking which can be present in the quantitative case. There are, in general, no simple analytical expressions like linear equations and their associated product-moment correlation coefficients for expressing relationships between items. Secondly, from zero-order

relationships, one cannot in general deduce higher order relationships very closely. The problems of prediction and correlation for qualitative data are quite different from that of quantitative data.[13]

The distinction between a variable and its frequency function. Confusion may arise as to what is a quantitative variable because when one is dealing with a statistical variable, one is always dealing with not *one* variable but with *two* variables simultaneously.

One variable is the content of interest; it is called *variate, statistical variable, random variable,* or *stochastic variable.* The other variable is the frequency function. The frequency function is always quantitative as it arises from a counting process. The variate itself may be either quantitative or qualitative; the fact that the frequency function is quantitative has nothing to do with the nature of the variate. Just because males and females can be counted does not imply that the variable sex is quantitative. It is this confusion between a variate and its frequency function that leads to improper phrases concerning "quantifying data." Votes for president are not put in a quantitative form when frequencies are recorded. Voting remains a qualitative act, even though it possesses a frequency function for a population.

It is undoubtedly such attempts to speak of a variate and its frequency function both in the same breath that has led to many verbal misunderstandings. (The presence from the very beginning of two variables is also the cause for difficulty in learning the logic of uncertain inference. The simplest problem in this field must involve at least two variables.)

The habit of least squares. Another of the difficulties in dealing with qualitative variates is a prevalent belief that qualitative classifications are in some sense "inferior" to quantitative classifications. This kind of belief will be found expressed even in the most recent textbooks. Attention has been paid, therefore, almost exclusively to the analysis of *quantitative* variates. Most of the work in elementary statistics can be said to be devoted to elementary aspects of least squares. The arithmetic mean is used extensively since it minimizes squares of deviations; the standard deviation is used to

[13] In his classical textbook, *An Introduction to the Theory of Statistics*, Yule devoted the first five chapters to studying the multivariate distribution of qualitative variables. It may be regarded as quite strange that statistical textbooks in the social sciences, where qualitative data are so important, have not followed suit, but instead fail to discuss material of this kind at all. A study of qualitative data as qualitative data could do much to help avoid the pitfalls of mechanically assuming that all data are quantitative and of mechanically using techniques and formulas developed and appropriate for only quantitative variables.

express dispersion; correlation indexes are defined in terms of ratios of variances; and so on. When *qualitative* variates do occur, they are examined ordinarily only from the point of view of how well they help predict quantitative variables, so that the focus again is on error in quantitative variates, measured in terms of squares of deviations. This is exemplified by the correlation ratio and the analysis of variance (including the t-test and related developments). Very little rigorous work has been done on the reverse problem of *predicting qualitative variates.*

The history of the prediction of qualitative variates, especially by sociologists, psychologists, and educational psychologists, reflects a habit of trying to cast everything into least squares, thereby losing sight of the data being dealt with. It has not been clearly recognized that the definition of *deviation* or *error* for qualitative variates must rest on its own feet and not be a mere analogy.

It has been pointed out elsewhere[14] that a principle appropriate for predicting qualitative variates is that of maximum probability, namely, minimizing the *number* of deviations. We cannot speak of *how far off* one qualitative value is from another when there is no metric to begin with, but we can always speak of the *number* of times we are off.

The habit of least squares must be abandoned if straightforward analysis of qualitative variates is to be made. Least squares may be useful as a mechanical tool later in the analysis—indeed some equations of scale analysis discussed in later chapters emerge from certain least squares considerations—but it is not the starting point nor guiding principle.

Since qualitative variates, like responses to attitude questions, are by definition not numerical, any technique which involves adding numbers does not apply to the variate values. Qualitative variates do not have arithmetic means, standard deviations, nor product-moment correlation coefficients. Qualitative variates do have frequency distributions, but the distributions cannot be analyzed in terms of most of the statistics used for quantitative variates. A qualitative variate does have a mode, namely, the variate value with the highest frequency. If a qualitative variable is to be predicted, the value used would be the modal value, because it minimizes the number of errors of prediction.[15] A prediction is either

[14] Louis Guttman, *loc. cit.*, especially pp. 258–263, 271–275. For an adaptation of this material, see J. P. Guilford, *Fundamental Statistics in Psychology and Education* (McGraw-Hill Book Co., Inc., New York, 1942), especially Chapter 10, pp. 176–197.

[15] See Guttman, *ibid.*, or Guilford, *ibid.*

right or wrong, so that predictability is measured simply by counting up the number of right and the number of wrong predictions. In quantitative variables, there is a numerical meaning to the size of error, which therefore permits computing things like averages and standard deviations; this is not the case for qualitative variables.

The danger of tetrachoric correlation coefficients. A widely used practice with respect to relating two dichotomies is the use of tetrachoric correlation coefficients. This is a very convenient method of estimating the product-moment correlation coefficient between two quantitative variates which are known to have a normal bivariate distribution. If there were no sampling error of people, the tetrachoric coefficient would be precisely equal to the product-moment coefficient for the population, and that is the only purpose in using the tetrachoric coefficient. It has no meaning apart from the product-moment coefficient it estimates.

If data are qualitative and are not obtained by dichotomizations of quantitative variables, the use of tetrachoric coefficients can be very dangerous and misleading. It is all too easy to take the position that if a question has two possible responses like "Yes" and "No," then these two responses should be regarded as artificial representations of some underlying quantitative variable which has a normal distribution. The important task the research worker should undertake is to *test* this assumption, instead of just accepting it uncritically. What is the variable of which this dichotomization is supposed to be a partition? How would one prove that such a variable exists? If the research worker cannot answer this question of how he would ever test such a hypothesis about an underlying variable, it would certainly not be safe to use the assumption.

Actually, scale analysis is a technique for testing the assumption that such a question is a dichotomy which can be regarded as a slicing of a quantitative variable. Instead of making assumptions and proceeding from there, scale analysis has as its aim the *proof or disproof* of the hypothesis. The specification of any particular type of distribution, like the normal distribution, is completely absent, however, from scale analysis. The existence of a single variable underlying a set of qualitative variables can be established or disproved without any reference to the nature of its frequency distribution. Only rank order is needed, and rank order does not depend upon the shape of the distribution.

Inconsistencies due to tetrachoric correlation coefficients. To illustrate the importance of testing the hypothesis that there is an under-

lying quantitative variable instead of merely accepting it, let us consider the following distribution of three dichotomies. Suppose we have three questions to which the possible answers are "Yes" and "No," and that for each question "Yes" indicates a "more favorable" attitude than "No." Let us suppose that these three questions are asked of a population and that only four kinds of people are found, with frequencies according to the following table:

Type of person	*Question 1*	*2*	*3*	*Frequency*
I	Yes	Yes	Yes	10
II	Yes	No	Yes	30
III	Yes	No	No	40
IV	No	No	Yes	20
				100

Out of 100 people, 10 said "Yes" to all three questions, 30 said "Yes" to the first and third questions but "No" to the second, etc.

If tetrachoric coefficients were to be computed between each of the possible pairs of these three variables, they will be found to be *completely contradictory*. There are three zero-order correlations possible: r_{12}, r_{13} and r_{23}. The table for r_{12} can be obtained from the preceding table as follows. The number of people who said "Yes" to both questions 1 and 2 is 10, namely all the people of type I. The number who said "Yes" to question 1 and "No" to question 2 is 70, the frequencies of both types II and III. 20 people, those of type IV, said "No" to both of the first two questions. No people at all said "No" to the first question and "Yes" to the second. The fourfold table of frequencies for questions 1 and 2 therefore looks as follows:

		Question 1 Yes	No
Question 2	Yes	10	0
	No	70	20

The tetrachoric coefficient for this table is +1, since there is a zero cell in the upper righthand corner. If these two dichotomies actually are slices of two quantitative variables which have a normal bivariate distribution, then those two quantitative variables must have a perfect, positive, product-moment correlation coefficient.

Let us now look also at the correlation between questions 1 and 3

and 2 and 3, respectively. The fourfold tables are computed from the trivariate table above and are as follows:

		Question 1	
		Yes	No
Question 3	Yes	40	20
	No	40	0

		Question 2	
		Yes	No
Question 3	Yes	10	50
	No	0	40

The tetrachoric correlation between questions 1 and 3 is -1, while between 2 and 3 it is $+1$. But these two correlations are impossible if r_{12} is $+1$ according to our first fourfold table. If two variables are perfectly correlated, the correlation of the first variable with a third variable must be exactly the same as the correlation of the second variable with the third variable. But here we have the paradox that the first two variables are perfectly correlated but have completely opposite relations with a third variable.

What the above example shows is that these three questions cannot be slicings of quantitative variables which have bivariate normal distributions with each other. The assumptions behind the use of the tetrachoric correlation coefficient cannot possibly hold here.

While the example just given is very extreme, it should serve to point out the possibility of inconsistencies creeping into analyses if tetrachoric correlation coefficients are used without any test being made that the assumptions behind them are satisfied by the data. More generally, it can be proved that inconsistencies can arise when the correlations are more moderate in size than $+1$ or -1. It is well known that the matrix of zero-order product-moment correlation coefficients between any quantitative variables must be Gramian. This is one of the basic conditions which make factor analysis in the sense of Spearman and Thurstone possible. However, if tetrachoric correlation coefficients are computed between dichotomies, then the resulting correlation matrix need not at all be Gramian. This is true even though the tetrachoric coefficients may be of the size ordinarily found for product-moment correlations in factor analysis problems. The reason is that zero-order relationships in general do not determine higher order relationships, and use of the tetrachoric involves only zero-order relationships.

In studying a multivariate distribution, which is the problem of scale analysis, *all* relationships must be studied if a proper descrip-

tion is to be obtained. Scale analysis studies all relationships simultaneously, not just zero-order relationships.

The measurement of error. Apart from algebraic inconsistencies that can arise from use of tetrachoric coefficients, perhaps an even more basic consideration is how to measure error of prediction or reproducibility. It so happens that in a perfect scale of dichotomous items all the fourfold tables will have a zero cell in the proper place so that all tetrachoric correlation coefficients are always equal to +1. In this sense, tetrachoric coefficients could be used to test the hypothesis that a perfect scale was present. If all the coefficients are not +1, then either the area is only approximately scalable, or is not scalable at all. But to test for a *perfect* scale, it is even easier to forget about any correlation coefficients at all and just to look and see if the proper cells have exactly zero frequencies.

The hypothesis of a perfect scale, it is safe to say, will almost inevitably be rejected in practice. There will almost inevitably be various kinds of errors present in empirical data. The practical problem is not to test whether a *perfect* scale is present, but rather whether an approximately perfect scale is present. What kind of error does an imperfect tetrachoric correlation imply?

Since the tetrachoric coefficient is equivalent to a product-moment coefficient when the underlying assumptions are fulfilled, this means that error is implicitly measured in terms of least squares. If the assumption of normality is fulfilled, then an imperfect tetrachoric coefficient implies that the squares of the errors of prediction of one quantitative variable from the other are greater than zero, and the coefficient deviates from perfection according as the squares of errors increase. Thus, the coefficient is concerned not with the predictability of one dichotomy from the other, but of one quantitative variable from another. The prediction of the actual dichotomies is more properly done by *counting the errors of prediction,* which is quite different from the least-squares prediction of the quantitative variable. For example, consider the following fourfold table between two "Yes"-"No" questions:

		Question 1	
		Yes	No
Question 2	Yes	30	10
	No	40	20

How would we actually predict question 2 from question 1 in practice? For this we must use the qualitative technique of modal probability.[16] If a person says "Yes" to question 1, we would predict him to say "No" to question 2; we would be right 40 times out of 70. If a person says "No" to question 1, we would again predict him to say "No" to question 2, being right 20 times out of 30. All told, we would be right 60 times out of 100. A similar process would be used to predict question 1 from question 2, with the total correct predictions being 70 out of 100.

In scale analysis, the problem, however, is not how well the items can be predicted from each other, but how well they can be predicted from the scale score. If the above two questions are from a scalable area, then they should ordinarily be highly reproducible from the scale scores, with at least 90 per cent reproducibility. There is only one nonscale type, that in the upper right-hand corner; and this has 10 per cent of the people. Scoring this type with either the "Yes-Yes" or "No-No" type will yield one error for each of these people. Hence there are ten errors out of two hundred responses, or 95 per cent reproducibility of the two items from themselves.

Of course, this reproducibility may be spuriously high; it will not necessarily hold up when more questions are added. Reproducibility is not to be determined in practice from only two questions. We use this little example here only to highlight the point that there is no necessary relationship between an indirect least-squares treatment of hypothetical quantitative variables, and a direct test of the reproducibility of the qualitative items themselves, when error is present. Scale analysis treats qualitative data *qua* qualitative data, and measures error accordingly. Tetrachoric coefficients do not deal with the items themselves, but bring in unnecessary (and untested) hypotheses which have no clear relationship to the problem of reproducing the qualitative data.

Biserial correlation coefficients. A correlation coefficient that comes closer than the tetrachoric to suiting the purpose of scale analysis is the biserial correlation coefficient. This coefficient can test the hypothesis that each item is perfectly reproducible from the scale score. In a perfect scale, the biserial coefficient of each item with the scale score is +1. But again, in the case of a *perfect* scale, there is no point to worrying about correlation coefficients at all.

[16] See Louis Guttman, *ibid.*

Just looking to see if the proper zero cells are present is all that is necessary.

The biserial coefficient falls down on two of the same counts that does the tetrachoric: its assumptions and its treatment of error. The biserial coefficient is used to relate a dichotomy to a given (continuous) quantitative variable. It assumes that the dichotomy is in reality but a slicing of another quantitative variable, and that the two quantitative variables—the given one and the hypothetical one—have a bivariate normal distribution. If these assumptions really hold, then the biserial coefficient is equal to the product-moment coefficient of correlation between the two quantitative variables.

Thus, like the tetrachoric coefficient, hypotheses about underlying quantitative variables and normal distributions are introduced. Since such hypotheses may or may not be true, they could be tested if the biserial correlations are to be used. But testing the hypotheses requires determining the hypothetical variables empirically. So we are back to the problem of scale analysis, of testing the hypothesis that all the items are slices of the same quantitative variable.

The biserial correlations will do this if the scale is perfect. Their interpretations become obscure if there is scale error, since they do not treat the items as qualitative data. There is no exact relationship in general between a biserial correlation and a coefficient of reproducibility. Treating the data on their own merits, as in scale analysis, is simpler and more direct, besides avoiding assumptions of normality and the like.

What correlation coefficients should be factored? If one desires to proceed mechanically to factor a set of qualitative items by using techniques that were designed instead for quantitative variables, it is necessary first to compute correlations between the items. What correlation coefficients shall be used? Several research workers have proceeded by using tetrachoric correlation coefficients. This tacitly means that the *items themselves are not being factored* but, rather, the hypothetical quantitative variables are. In order to prove that this procedure is proper even for hypothetical variables, it is necessary to prove that the assumptions underlying the use of the tetrachoric coefficients are justified. It is possible to disprove this assumption in some cases by proving that the matrix of tetrachoric intercorrelations is not Gramian. It is not possible to do the converse, however, to prove that the assumptions are correct by showing that the matrix is Gramian. The matrix can be Gramian

and the assumption still be false. The only way really to prove the assumption is to exhibit empirically the underlying quantitative variable for each of the items. If the research worker has not done this, then his use of tetrachoric correlation coefficients must be held in doubt.

If the assumptions underlying the tetrachoric correlation coefficients are erroneous, then of course a Spearman-Thurstone factor analysis will be misleading even with respect to hypothetical quantitative variables. From a scale analysis, it can be predicted quite well what the factor analysis will show. If the items themselves have high reproducibility, then a factor analysis of tetrachorics will show, not *one* common factor, but *several* common factors. The tetrachorics will ordinarily not follow the single factor pattern because they are not affected uniformly by scale error. Scale error in a dichotomy with a 90–10 split will ordinarily reflect differently, in a tetrachoric, from scale error in a 50–50 split. Items with similar marginals will tend to have higher intercorrelations than items with dissimilar marginals. Hence, if the sample of items happens to have several items in one range of marginals, and several items in another range, it might be expected that scale error will cause the tetrachorics to reveal *two* common factors algebraically in the Spearman-Thurstone sense. The tetrachorics will reflect not just the single scale factor, but all sorts of artifacts, since such coefficients are quite sensitive to scale error and to extraneous considerations like marginal frequencies.

Other research workers have avoided the use of tetrachoric coefficients and instead have used point correlation coefficients as a basis for a factor analysis. The point correlation coefficient is a product-moment coefficient so that the matrix of zero-order point correlation coefficients is always Gramian. It still does not follow, however, that a factor analysis of point intercorrelations will properly describe the multivariate distribution of the attributes. Indeed, we shall now show that if the items form a perfect scale, so that each item is a simple function or cut of the same quantitative variable, then the point intercorrelations will be completely misleading. If there is but a single factor in the sense of scale analysis —that is, if each of the items can be reproduced from but a single quantitative variable—then a Spearman-Thurstone factor analysis of point correlations will never show this to be the case, but will instead exhibit apparently many factors. For convenience, let us consider four dichotomous questions. This time we consider the

four questions to form a scale so that there are five perfect scale types as shown in the following table:

	Question				
Type of person	*1*	*2*	*3*	*4*	*Frequency*
I	Yes	Yes	Yes	Yes	10
II	Yes	Yes	Yes	No	20
III	Yes	Yes	No	No	40
IV	Yes	No	No	No	20
V	No	No	No	No	10
					100

The multivariate distribution of the questions as shown in the table follows that of a scale pattern. The fourfold tables for the zero-order correlations can be determined as indicated in the previous example. The point correlation coefficients computed from the fourfold tables are as follows:

POINT INTERCORRELATIONS OF FOUR SCALE DICHOTOMIES

Question	*Question*			
	1	*2*	*3*	*4*
1	—	.51	.22	.11
2	.51	—	.43	.22
3	.22	.43	—	.51
4	.11	.22	.51	—

In order for there to be a single common factor in the sense of Spearman, the tetrads must vanish. This vanishing occurs for only one kind of tetrad in the above table, but not for another. A vanishing tetrad is in the upper right-hand corner, consisting of the elements in rows 1 and 2, and columns 3 and 4: $(.22)(.22) - (.43)(.11) = 0$. A nonvanishing tetrad consists of the four elements in rows 2 and 3 and columns 1 and 4: $(.51)(.51) - (.22)(.22) = .21$. In general, the following is true for the matrix of point intercorrelations of scalable dichotomies, where the questions are arranged in order of their marginals: all tetrads involving elements on only one side of the diagonal will vanish, but all tetrads involving elements from opposite sides of the diagonal will not vanish. Since there must be nonvanishing tetrads in general, Spearman's criterion is not satisfied by a scale.

This can be shown for the general case as follows. From the fact that there is a zero cell in each fourfold table from a scale, the formula for the point correlation simplifies as follows. If p_j is the

proportion of people who give a "favorable" answer to the jth item and p_k is the proportion who give a "favorable" answer to the kth item, and if p_j is less than p_k, then the point correlation between the two items is

$$r_{jk} = \sqrt{\left(\frac{p_j}{1 - p_j}\right)\left(\frac{1 - p_k}{p_k}\right)} \qquad p_j \leqq p_k$$

If we let $b_j = \dfrac{p_j}{1 - p_j}$ and $b_k = \dfrac{p_k}{1 - p_k}$, then the formula for the point correlation becomes

$$r_{jk} = \sqrt{\frac{b_j}{b_k}} \qquad b_j \leqq b_k$$

It will be recalled that in Spearman's single factor theory, if a_j and a_k are the respective correlations of the jth and kth quantitative variables with the common factor, then the correlation between the two variables is

$$r_{jk} = a_j a_k \qquad (j \neq k)$$

Comparing these two formulas for the correlation coefficient, it is clear that they are contradictory. If there is a single factor in the sense of scale analysis for qualitative data, there cannot in general be a single common factor in the Spearman sense, so that more than one apparent common factor will emerge from a Spearman-Thurstone analysis.

The components in a scale. If a set of dichotomies forms a scale, it is clear then that the Spearman-Thurstone factor analysis will not detect this fact. On the other hand, a scale analysis can tell us what to expect from a factor analysis and shows us in what respects the factor analysis will be misleading. The equations of scale analysis as analyzed in Chapter 9 show further why a factor analysis using point correlations apparently discovers more than one factor, whereas in fact there is but a single variable present in the sense of scale analysis. These equations show that a scale of dichotomies can be resolved into many components. Each item can be regarded as having two quantitative values, zero and unity, which can be expressed as a linear function of these components. In a certain sense, the first component is essentially the scale scores; the second component has been identified as the intensity function; the third and remaining functions have not yet been named, but it is known that

they are of an oscillatory character. The important feature of these functions is that *they are all perfectly related*, albeit in a curvilinear fashion, *with the scale scores*. The linear correlations of these functions with each other, including their correlations with the scale score, are all zero, yet they all are perfect curvilinear functions of the scale score. It is this paradox of having perfect curvilinear functions which are, nevertheless, linearly uncorrelated that throws factor analysis off. The Spearman-Thurstone approach to factor analysis is a completely linear one and is not adequate for analyzing the curvilinearities which are inherent in the scale pattern. The fact that a scale has a single factor but nevertheless many components which are perfect functions of that factor makes a Spearman-Thurstone factor analysis inappropriate here.

One of the interesting features of scale analysis is that, while it has the implications of these oscillatory components, nevertheless there is no need to worry about them in testing the hypothesis of scalability. The innocent-looking scale pattern is all that has to be analyzed by any of the several simple scale analysis techniques now available. In doing this, the mathematics of scale analysis are implicitly employed, but there is no need for the clerk to know anything about them. Factor analysis with its heavy computations is a vastly more tedious and expensive kind of procedure. Its complexity, however, does not mean therefore that it is appropriate for qualitative data. Factor analysis may be a proper approach for some problems involving quantitative variables. As for qualitative data, it is theoretically more desirable and in practice far simpler to analyze qualitative data *qua* qualitative data.

Communalities and reproducibility. It may be worth pointing out some further differences between factor analysis and scale analysis with respect to how they intend to study the structure of their respective kinds of variables. Scale analysis tests a single factor hypothesis for qualitative data: can a single quantitative variable suffice to reproduce the response of each individual to each question? The efficiency of this reproducibility is measured by the simple per cent of correct estimates of persons' responses from their scale scores.

In the Spearman-Thurstone approach to factor analysis, even in the single factor theory it is not attempted to reproduce each person's score on each quantitative variable from his score on the common factor. Actually, there is no single factor theory in the Spearman-Thurstone approach because of the distinction made

between common factors and unique factors. Each variable in a set is thought to be composed of not just the common factors, but also of a unique factor (which may be a composite of several specific and error factors). In order to reproduce a person's score, then, on one of the original variables, one would have to know not only his common factor score but also his unique factor score. Since each variable has a unique factor, in general there are $n + 1$ or more factors implicit in a factor analysis of n variables, and not just one factor. The common factor does not account for the whole variance of each variable, but only for part of the variance, which is called the communality. The purpose of the analysis is to find out if common factors can be found which will reproduce the *correlation coefficients* between the variables. The number of common factors does not depend upon the size of the communalities to any appreciable extent. The communalities can be very low, yet one common factor may be found or many common factors may be found; or the communalities may be high, yet either one common factor may be found or many common factors. The correlations can be reproduced perfectly even though the individual scores can be reproduced only very inaccurately from the common factors. Thus, a single common factor in the sense of Spearman and Thurstone for quantitative variables does not imply the same kind of reproducibility for its variables as the single common factor of scale analysis implies for its qualitative items.

As a final comment, it might be pointed out that in the mental test field, where factor analysis originated and has been most widely used, the quantitative variables employed are themselves scores obtained from sets of qualitative items. An achievement test is a sample from a universe of qualitative data. It might be suggested, therefore, that before a factor analysis is applied to scores derived from sets of items, each test separately be first subjected to a scale analysis. If the test is not a scale (or quasi scale), then it should not be represented by only a single score in the factor analysis but perhaps should be broken down into subareas, each of which might prove to be scales or quasi scales. If tests used in a factor analysis are neither scales nor quasi scales, then it is hard to see how a factor analysis can show any clear-cut results. Perhaps one reason for low communalities is the fact that the tests used are not scales or quasi scales, so that the scores on them do not have much internal meaning.

Generalized factor analysis: latent structure theory. The develop-

ment of the theory of latent structure for qualitative data by Lazarsfeld, which was also an outgrowth of the work of the Research Branch, enables us now to view factor analysis in a broader setting.

The generalized theory of common factors—or of latency—as developed by Lazarsfeld, can be stated as follows. Consider a set of n manifest variables $x_1, x_2, \cdots, x_n$ whose structure is to be analyzed for a given population. These variables may be all qualitative, partly qualitative and partly quantitative, or all quantitative. That is, we are not at all limiting the forms of the variables. Some may be dichotomies, some trichotomies, some may have ten categories with no particular order to them, some may have many ordered categories, some may be discrete quantitative variables, some may be continuous quantitative variables.

The basic hypothesis of the generalized theory is that the population can be divided into a number of subpopulations, such that *within* each subpopulation the n manifest variables are *statistically independent*. Each subpopulation can be called a *latent category*. The set of latent categories can be called the *latent set*. The latent set may be an attribute—that is, consist of nonnumerical and even nonordered categories; if in addition the manifest variables are also attributes, then we have Lazarsfeld's theory of latent structure. The latent set may be a single quantitative variable; if in addition the manifest variables are quantitative (and if zero correlations also imply complete statistical independence), then we have Spearman's case of a single common factor. The latent set may comprise more than a single quantitative variable—that is, be a vector variable; if in addition the manifest variables are quantitative (and if zero correlations also imply complete statistical independence), then we have Thurstone's case of multiple common factors.

All this can be stated analytically in terms of the frequency distributions involved. Denote the latent set by λ. If there are m subpopulations involved all told, then λ has m categories or values. The multiple frequency function of the n manifest variables for a given value of λ can be denoted by the usual notation for conditional frequency functions: $f(x_1, x_2, \cdots, x_n \mid \lambda)$. If none of the manifest variables involved is continuous, then each value of f is an actual probability—such as in Lazarsfeld's case. Otherwise, f represents probability *density* as is usual for continuous variates.

Similarly, we can denote the distribution of each manifest variable separately for a given value of λ by the notation of conditional distributions. The n conditional frequency functions are: $f_1(x_1 \mid \lambda)$,

$f_2(x_2 \mid \lambda), \cdots, f_n(x_n \mid \lambda)$. Again, an f_j denotes either an actual probability or a probability density, according as the distribution is discrete or continuous.

The basic hypothesis of statistical independence within subpopulations is then equivalent to the statement that the multiple conditional distribution is equal to the product of all the first order conditional distributions:[17]

$$(1) \qquad f(x_1, x_2, \cdots, x_n \mid \lambda) = f_1(x_1 \mid \lambda) f_2(x_2 \mid \lambda) \cdots f_n(x_n \mid \lambda)$$

Equation 1 contains both the Spearman-Thurstone structures and the Lazarsfeld structures as well as possible mixtures of the two. As an algebraically degenerate case, it also contains the perfect scale. It is the basis of what Lazarsfeld calls "accounting equations."

For the perfect scale, the latent set λ is simply the scale ranks or scores. The conditional frequency distributions are all degenerate in that they have only two probability values possible: unity and zero. All people with the same rank have the same response on each manifest variable; they are all in but one category of each x_j, for which f_j is then unity, and the remaining categories have their f_j's vanish. Furthermore, the nonvanishing f_j form the parallelogram pattern of the scalogram. The scale types are defined in the left member of equation 1 by those values of the manifest variables for which f is unity. The nonscale types do not occur; for them f is zero.

As Lazarsfeld has shown, the quasi-scale pattern may also be viewed as a special case of equation 1. The set λ is the ranks or scores on the quasi scale. This time, however, the conditional distributions are not degenerate; they have probabilities between zero and unity. Furthermore, the categories of the manifest attributes can be arranged in such an order that the conditional probabilities form a gradient within each manifest attribute. Thus, a person in a higher category of λ has a higher probability of being in a higher category of x_j than a person in a lower category of λ, and this is true for all x_j.

The perfect scale requires the discrete probabilities of zero and unity; it therefore implies that the manifest variables are discrete. Indeed, the practical work to date has been with attributes which are discrete manifestations. On the other hand, although quasi

[17] This is equation 1.1 in Chapter 10.

scales are also discussed in these chapters from the point of view of qualitative data, the concept holds even for continuous quantitative variables, including Spearman's single common factor.

If each of the manifest x_j and also the latent λ are continuous quantitative variables, and if they have a *normal* multivariate frequency distribution, then the condition of independence of equation 1 is equivalent to Spearman's relationships between correlation coefficients:

$$r_{x_j x_k \cdot \lambda} = 0 \quad (j \neq k) \tag{2}$$

Equation 2 states that the correlation between any two manifest variables is zero within each subpopulation. This leads to Spearman's tetrad equations and the related algebra. Thus, for this quantitative case, an algebra of correlation coefficients suffices to study equation 1.

That Spearman's single common factor pattern is also a quasi scale can be seen from the regressions of the manifest variables on the latent variable or factor. The regressions are linear, so that a person with a higher value of λ is more likely to have a higher score on x_j than a person with a lower value of λ, and this is true for all the manifest x_j. (The case of negative regression coefficients can always be taken care of by rearranging the appropriate manifest values, which in this case means simply reversing the signs of the x_j that require it.)

Generalized factor analysis and latent structures. In Chapters 10 and 11 below, Lazarsfeld develops the theory of latent structure where both the manifest variables and the latent set are qualitative. The structure of the *latent dichotomy* is the case where λ has only two categories, and more generally λ may have m categories with no particular order among themselves. The subpopulations may differ in kind rather than in degree as in the case of the perfect scale or quasi scale. No assumptions of rank order of any kind are required for the population.

For Lazarsfeld's qualitative structures, the probabilities are all discrete. Each member of equation 1 consists of actual probabilities. Chapters 10 and 11 show how to study these probabilities by means of an algebra of matrices of frequency distributions. These matrices do not exist when the manifest variables are continuous, for then we have probability densities or ordinates, rather than actual discrete probabilities, in equation 1. An appropriate algebra

for the continuous case—if the multivariate distribution is normal —is that of Thurstone's multiple common factors.

We have already seen how Spearman's structure is a quasi scale for continuous variates. If Spearman's hypothesis is not satisfied, then the categories of λ do not form a single quantitative variable. Thurstone's techniques are concerned with equation 2 where λ is a *set* of quantitative variables. This leads to a study of the matrix of correlation coefficients—not to be confused with Lazarsfeld's matrices of probabilities—which is equivalent to a study of equation 1 when the multivariate distribution is normal (so that zero correlations imply statistical independence).

Thus we see again how the algebra required for qualitative data differs from that for quantitative (especially continuous) variables.

On the other hand, from equation 1 we can see that it is possible always to apply the qualitative techniques to quantitative data! It is always possible to divide quantitative manifest variables into class intervals, even into dichotomies, without disturbing the structure of equation 1. Or, for qualitative manifest data, it is possible to combine categories in any way one pleases without disturbing the form of equation 1. This theorem follows immediately by observing that summation (for discrete categories) or integration (for continuous categories) over values of an x_j on the right of equation 1 merely replaces the f_j's of those values by the conditional probability of the combined categories or of the interval of integration. The multiplication of probabilities on the right of equation 1 is undisturbed—so that *exactly the same latent set holds for the combined data as for the original data.*

It is interesting that the property of invariance of structure under combination of categories holds not only for scales and quasi scales but for all latent and common factor structures.

Dividing continuous manifest variables into intervals yields discrete probabilities, and these can be studied by Lazarsfeld's algebra. Of course, no different latent categories can emerge that were not present in the original data. On the other hand, latent categories can be lost (through implicit combination) by the combination of manifest data. In the case of scales, we know that scale types are lost by combining categories. In the case of quantitative variables, it may be that whole common factors can be lost. In any event, it may well be worth exploring to what extent it is worth while to use qualitative techniques on quantitative data. At present, there is no method for nonnormally distributed quantitative variables; the

Spearman-Thurstone "factors" do not actually satisfy equation 1 unless zero correlations imply full statistical independence. The nonmetric technique of, say, dichotomizing the variables and then using Lazarsfeld's algebra does provide a method for such a case. Since common factors may be lost by dichotomization, it may be that a series of analyses, based on different kinds of dichotomizations of the manifest variables, could tease out all the latent categories—if a latent structure holds at all.

Justification for the combination of categories. One of the interesting and important consequences of equation 1 is the justification which it provides for a practical procedure used in scalogram analysis. If manifest items have three or more categories, it has very often been found necessary in practice to combine some of these categories to reduce scale error. The theorem of the previous section shows what the effect of such an operation is if the uncombined data have a latent structure. The scale defined by the combined data *must* be part of the original latent structure; no latent types are lost. Each scale type that emerges is also a latent category.

What is lost by combining manifest categories is some possible latent categories that reflect factors additional to the scale variable. Such factors can be nonessential to the study and hence ignored, such as a factor of verbal habit described before.

On the other hand, all that may be lost is but a difference between two quasi-scale types when two adjacent categories are combined and then found to fit a perfect scale. The categories may each separately be found to satisfy equation 1, but to have scale error, whereas their combination (which then must also satisfy the equation by our theorem) will have no scale error. This may imply a "quasi-scalish" relationship between particular categories, though the perfect scale pattern holds for the data as a whole.

Other possible structure theories. We now wish to point out three things which are *not* necessarily to be inferred from the foregoing discussion. First, the case of the perfect scale is not necessarily to be regarded as only a part of the theory of latent structure. Second, latent structure is not to be regarded as the only kind of structure theory possible for a universe of content. And third, the quasi-scale pattern as seen from the point of view of latent structure may not be the only type of quasi scale.

That the perfect scale satisfies equation 1 is not surprising. In fact this is *mathematically trivial.* This triviality is not just a consequence of the degenerate probabilities involved, but of an obvious

corollary of equation 1. It is obvious that, if the set of *manifest* types is used for λ, then equation 1 is satisfied no matter what the original frequency distribution is. That is, equation 1 is always satisfied mathematically if the manifest set is used for the latent set.

In the perfect scale, there is a one-to-one correspondence between scale types and ranks, so that using ranks for λ is equivalent to using the manifest types. This is trivial as far as equation 1 is concerned. In the same way, for example, nonscalable data might have any values assigned each type, and these values used for λ; this will always satisfy equation 1, no matter what the structure of the data. Or again, for quantitative variables, any nonsingular linear transformation of the n manifest variables can be used for λ (including the original x_j themselves), and equation 1 will always be formally satisfied no matter what the manifest variables are.

Therefore, the importance of the scale pattern does not lie in that it satisfies equation 1. Its importance and its properties lie in the *internal pattern of the scale itself*, which has nothing whatsoever to do with satisfying equation 1. Equation 1 says nothing about ranking individuals or similar problems. It merely gives a definition of latency, and of latent categories. It so happens that for the perfect scale, the latent categories *are always manifest*, so that the notion of latency adds nothing at all; it actually contributes nothing to the mathematics of scale theory (as, for example, developed in Chapter 9). The theory of order of scale analysis is thus quite independent of latent structure theory. The relationship between the two is mathematically trivial. Indeed, scale theory has a mathematical relationship to another theory of structure that is radically different from that of equation 1.

On the other hand, the fact that *quasi* scales can be viewed from the perspective of equation 1 is far from trivial. Here, the latent types and the manifest types do not coincide. Indeed, latent structure theory is important only for those cases where the number of latent types is less than the number of manifest types. This condition is satisfied by quasi scales, whereas it is not by perfect scales. It is true that the problem of order within quasi scales is not touched on by equation 1, but requires additional considerations; however, equation 1 introduces the latent categories upon which to base the additional considerations.

There is a different theory of structure from that of Spearman-Thurstone-Lazarsfeld for which the perfect scale is a special case,

and *nontrivially* so. This theory does not require any notion of latency or common factors, but is entirely in terms of manifest variables. It may be called a theory of *conditional order*. Its basic hypothesis is that the manifest variables can be arranged in such an order that, if one variable is held constant, then the *immediately* preceding and succeeding ones are statistically independent. This hypothesis is expressed by the formula:

$$(3) \qquad f(x_i, x_k \mid x_j) = f_{ij}(x_i \mid x_j) f_{kj}(x_k \mid x_j) \qquad (i < j < k)$$

It is not intended to enlarge on this theory here but only to point out that the scale parallelogram satisfies equation 3 nontrivially (albeit degenerately). Furthermore, if the x_j are all quantitative and satisfy equation 3, then their correlation matrix can be proved to have principal components that satisfy the difference equations for scale weights of Chapter 9. Thus, the *weight* equations of Chapter 9 apply also to a general case of equation 3. The scale of dichotomies is the special case where the *score* equations also have the oscillatory properties proved in Chapter 9.

The structural equation 3 in general is mutually exclusive of equation 1, and it has a general mathematics of order that applies in particular to the perfect scale.

It remains to be seen whether other general structural equations besides those of equations 1 and 3 can be found that will also lead to alternative useful theories of quasi scales. There is still a great amount of work to be done on the general theory of gradient in responses. Latent structure theory deals with the gradient in terms of a gradient within latent categories. But there may still be an obvious gradient in manifest responses and yet the hypothesis of latent distance need not be satisfied. That is, subpopulations may be defined and ordered so that a manifest gradient exists, and yet statistical independence need not hold within these subpopulations. This is a problem that needs further research.

CHAPTER 7

THE INTENSITY COMPONENT IN ATTITUDE AND OPINION RESEARCH[1]

SCALE analysis provides an answer to the problem of how to rank people from high to low with respect to the particular area being studied. From the rank order of a person it is possible to reproduce his answers to all the questions. In the case of dichotomies, it is possible to predict that a person who endorses a more extreme statement must endorse all less extreme statements. More generally, persons who endorse one statement must all have higher ranks than persons who do not endorse the same statement. Scale analysis thus provides a rigorous definition of a unidimensional scale continuum.

Given such a rank order, we can say that *A* is *more* favorable than *B* who is *more* favorable than *C* and so forth. But what can we say about *A*, *B*, and *C* with regard to whether or not they have a "favorable" or an "unfavorable" attitude? One person can be more favorable than another, yet both may in some sense be "unfavorable." Can we fix any cutting point along the scale continuum which would permit the division of the population into "favorable" and "unfavorable," rather than only a relative ranking in terms of degrees of favorableness? Are the two ends of a scale continuum *opposite* in any sense?

This problem of the meaning of the different extremes of a scale continuum is clearly recognized by McNemar, who states, "Another aspect of the single continuum problem is the meaning of the two extremes of a scale. If one end indicates a favorable attitude, does the other end represent unfavorableness or something else? For instance, is Fascism the opposite of Communism? . . . Are we dealing here with one or two dimensions? To interpret the extremes as being real opposites or to interpret the extremes as on different con-

[1] By Edward A. Suchman.

tinua is precarious in the absence of knowledge concerning the real meaning of the extremes." [2]

The Problem of a Cutting Point

The need for an objective definition of a cutting point is especially vital for public opinion polls. Public opinion analysts and other workers in the field of attitude measurement have long been aware of the problem of question "bias." There are numerous cases on record in which differently worded questions dealing with the same issue have produced different percentages of the population apparently "opposed to" or "in favor of" the issue being studied. Slight changes in the question wording, in the order of presentation of the answer categories, in the position of the question in the questionnaire, and many other factors known to all pollsters may affect the findings of one's poll. As a result of this variation in polling results, a great deal of attention has been given to the determination of criteria to be used in constructing an "unbiased" question. For the most part the criteria suggested in the past have been subjective, being based upon "expert opinion."

How can we account for this variation in poll results depending upon the question asked? As was discussed at length in the previous chapters, the answer we propose involves the concept of an attitude or opinion universe. Any single question asked on an issue is but one sample of all possible questions that could be asked on that same issue, and the proportion saying "Agree" or "Yes" to these questions can theoretically range from 0 to 100 per cent.[3] For example, a survey of soldiers' attitudes toward the British asked, "Do you agree or disagree with the following statements about the British?"

______ The British are doing as good a job as possible of fighting the war, everything considered.

______ The British always try to get other countries to do their fighting for them.

We find 80 per cent of the soldiers *agreeing* to the first statement, but only 48 per cent disagreeing to the second statement. Which of

[2] Quinn McNemar, "Opinion-Attitude Methodology," *Psychological Bulletin*, Vol. 43, No. 4 (July 1946), p. 299.

[3] Compare the "tree" of public opinion in Hadley Cantril (editor), *Gauging Public Opinion* (Princeton University Press, Princeton, N.J., 1944), pp. 26–27. "Interventionist" sentiment was found to vary from 8 per cent to 78 per cent depending upon the particular question asked.

these questions is an "unbiased" question with respect to the entire issue? How many soldiers are "favorable" toward the British?

It should be clear, then, that it is difficult to conceive of a marginal frequency distribution which can adequately represent the whole universe when obtained from but a single question. There are innumerable nuances that can be put into questions by different shades of wording, different check lists, different ordering of questions in the questionnaire, different atmospheres of answering a questionnaire, and the like. Each question used in a survey is but a sample of all the variations of it that could have been used instead.

From these variations in polling results depending upon question wording, and from lack of "expert" agreement as to what is "biased" or "unbiased," it appears quite clear that what is greatly needed is some *objective* method of dividing the respondents into the *same* proportions pro and con *regardless of question wording.* With such a method different opinion pollsters would come out with the same results, *even though they should ask different questions* concerning the same topic. These results would be independent of many of the various "biases" to which public opinion polling today is subject.

Is such a method possible? Can there be an *objective* method for determining an *invariant* dichotomization of the population with regard to an opinion or attitude area?

The Intensity Function

An answer to the above question must involve the determination of a cutting point that is "meaningful" with respect to a whole scale continuum. A suggested method for doing this is proposed in the following sections. This method utilizes what we call the *intensity function.* It is based upon the concept of a scalable area in which individuals are ranked from high to low and for which it is possible to measure the *intensity of feeling* with which people at different scale positions hold their attitudes or opinions. Intensity of feeling is conceived to be strongest at both ends of the scale continuum and to decrease as one moves toward the middle. The scale position with lowest intensity serves to mark off a point of "neutrality" which provides a division of the population into two groups, one positive and the other negative. Furthermore, as a consequence of the scalability of the area, this division does not depend upon the particular sample of questions used. The cutting point is determined objectively and is invariant with respect to wording of questions.

In the following sections we will first discuss in detail the theoretical foundations for the intensity function. Then we will report upon various techniques used in practice to measure the intensity function and show several examples of the actual determination of cutting points by means of the intensity function. Finally, we will present two examples of the independence of the intensity function and the cutting point from the "biases" involved in any specific questions asked.

The Basis of the Intensity Function

We conceive of a content scale which ranks people from high to low on a single content continuum. We furthermore conceive of an intensity scale which ranks people from strong to weak on a single intensity continuum. Correlation of these two scales—content and intensity—produces a *J*- or *U*-shaped curve. As one moves from one end of the content scale, intensity of feeling decreases until a point is reached at which intensity of feeling begins to increase again. This point represents the content position of persons with lowest intensity. It may be thought of as a point of indifference or neutrality.[4]

We shall call this point the zero point. This zero point appears to offer us an invariant cutting point on a scale running from favorable to unfavorable which permits us to divide the population into two groups which can be meaningfully labeled as "positive" and "negative." Since the rank order of people on *both* content and intensity is independent of any particular sample of questions asked, the shape of the *U* curve and the location of the lowest point on this curve also remains independent of any particular sample of questions asked. The intensity curve is fixed for any single attitude universe—any sample of questions from that attitude universe will produce the same results.

The Ideal Curve

An example of an ideal intensity curve with a perfect correlation between content scores and intensity scores is given in Table 1a, assuming a sample of content consisting of ten equally frequent

[4] The results of several previous studies on intensity of feeling seem to bear out this rational explanation for the existence of a *U*-shaped intensity curve. For example, Katz shows that the more "extreme" an attitude is, the more intensely it is likely to be held. D. Katz, "The Measurement of Intensity," in Cantril, *ibid.*, Chapter 3. This chapter deals with several different methods of measuring intensity of feeling in opinion surveys.

ranks and a sample of intensity consisting of six ranks. Along the horizontal axis is plotted the percentile order of respondents according to their content scale scores, while along the vertical axis is plotted the percentile order of individuals according to their intensity scores. Theoretically, the intensity scores must have a perfect curvilinear relation with the content scale scores. All persons in any one rank or content must have the same rank on intensity.

TABLE 1a

PERFECT CORRELATION TABLE OF CONTENT SCORES BY INTENSITY SCORES

Intensity rank	*Content rank* (*Neg.*) *0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	(*Pos.*) *8*	*9*	*Total freq.*	*Cum. freq.*
5 (High)										*10*	10	100
4	*10*								*10*		20	90
3		*10*						*10*			20	70
2			*10*				*10*				20	50
1				*10*		*10*					20	30
0 (Low)					*10*						10	10
Total freq.	10	10	10	10	10	10	10	10	10	10	100	
Cum. freq.	10	20	30	40	50	60	70	80	90	100		

At least one of the extreme content ranks must correspond to the highest intensity rank, while the arms of the *J* or *U* must come down to the lowest intensity rank.

In the intensity curve there are usually two content ranks for each intensity rank, except for the lowest intensity rank. The content percentile of the persons with the lowest intensity rank therefore corresponds to the zero point. Were it not for the presence of errors in the empirical measurement of the intensity function—a problem which will be discussed later—it would be possible to determine the zero point along an attitude continuum simply by determining the content rank for persons with the lowest intensity scores. This means locating a point along the content continuum which contains the persons who are most "indifferent" or "neutral."

The zero point thus established has a definite meaning with respect to the attitude universe. It separates those whose intensity of feeling goes up as their attitude is *more* favorable, from those

whose intensity of feeling goes up as their attitude is *less* favorable. A further important property is the invariance of the zero point. The intensity curve for one sample of items is essentially the same as the intensity curve for any other sample of items, and is, in fact, essentially the intensity curve for the entire scalable universe. The basic importance of the concept of testing a sample of items to determine if they belong to a scalable universe has been discussed in the previous chapters on scale analysis. It is because two different samples belong to the same scalable universe that they produce the same rank orders of individuals on both content and intensity, and therefore result in the same intensity curve and zero point. Two investigators would have to come out with the same percentages of the population on either side of the zero point regardless of the questions asked, provided only that they dealt with the same scalable issue. An empirical demonstration of this property will be given later.

The Problem of Error

The research problem is to measure intensity empirically. If this can be done, the zero point can be determined by plotting intensity of feeling against degree of favorableness. The ideal curve is probably impossible of attainment in practice. Present techniques for the measurement of intensity all contain so much error that in all cases intensity has been measured as a "quasi scale." [5] One essential difference between a scale and a quasi scale is that from scale scores one can reproduce the response of each individual to each item (within the margin of error provided by the coefficient of reproducibility of the scale), while in a quasi scale this internal reproducibility is not possible. However, the rank order of respondents on a sample of items from a quasi-scale area would still be essentially the same rank order as that in the universe of all items that could have been used. Therefore, the ordering of the persons on content and on intensity is essentially invariant with respect to the particular sample of questions used.

The fact that intensity forms a quasi scale shows that there is one dominant factor in the sample of intensity questions asked, which we call intensity of feeling, and many other smaller factors. It is the presence of this dominant factor which permits the appear-

---

[5] For a more complete discussion of a "quasi scale" see Chapters 5 and 6.

ance of the desired *U* curve. Because of the error in our present techniques of measuring intensity, it is necessary to use some average measurement of the intensity scores. The regression medians of intensity on content do yield a *J*- or *U*-shaped regression curve and seem to afford an adequate working picture of the intensity function. It is necessary to use the entire population sample for the determination of the intensity component. *Proportions* of the population at any given rank or on one side of the zero point are subject to ordinary sampling error.

It is also advisable in an intensity analysis to have as many scale ranks as possible for both content and intensity scales. The more points one has to plot, the more detailed will the intensity curve become. In particular, the close determination of a zero point requires questions whose cutting points are in the region of the bending of the curve, as contrasted to the determination of rank order alone which requires only the number of questions needed for the desired number of rank orders.

It is to be expected that as further research develops better techniques for the measurement of intensity of feeling, the obtained intensity curve will more closely approximate the ideal curve.

A Technique for Observing Intensity

A simple approximation of the intensity function has been successfully attained by asking a question about intensity of feeling after each content question. One form used for an intensity question is simply: "How strongly do you feel about this?" with answer categories of "Very strongly," "Fairly strongly," and "Not so strongly." Repeating such a question after each content question yields a series of intensity answers. Using the same procedure as was outlined previously for the content answers, these are scored and each respondent is given an intensity score. The intensity scores are then cross tabulated with the content scores.

This method of measuring intensity of feeling corresponds to the technique of securing a judgment of a "whole" by asking for judgments of different "parts" of the "whole." [6] It is often very difficult to determine judgments of a complex object, but breaking that object down into several simpler components which can be judged

[6] This approach is developed in greater detail in P. Lazarsfeld and W. Robinson, "The Quantification of Case Studies," *Journal of Applied Psychology*, Vol. 24, No. 6 (December 1940), pp. 817–825.

may permit one to determine a judgment of the complex object by combining the judgments of its component parts. Thus, for example, one could determine people's judgments of a horse not by asking the person whether he liked the horse as a whole, but by asking for simple "yes"-"no" judgments about different parts of the horse, such as his head, his neck, his feet and his tail, and then combining these separate judgments. In a similar way, while it might be difficult to determine how strongly people felt about some broad area, such as their attitude toward their officers, asking them how strongly they felt about various specific aspects of their officers' behavior permits one to combine these "smaller" judgments into a general judgment of intensity of feeling on the area as a whole.

The cross tabulation of intensity scores by content scores should produce, in an ideal case, a table similar to Table 1a in the previous section. However, due to the presence of error, we find that persons with the same content score may have different intensity scores. We therefore compute the median intensity for each content rank and use the curve of medians as an approximation of the true intensity function. To locate each median we determine for each content rank the individual who occupies the median position on intensity for that content rank and then find out what his intensity percentile is in *the entire population.* This is done by the ordinary formula for interpolation for the median of a continuous distribution when only class intervals are given. Medians are used, rather than arithmetic means or similar averages, because medians are independent of any metric apart from rank order.

A proper graphic presentation of the curve of medians is to express the data not in the crude ranks observed, but in estimates of what the rank on the whole attitude universe would be if indefinitely many questions had been asked. The percentile metric offers a way of doing this. The cumulative frequencies for the row (content) and column (intensity) marginals are computed at the bottom and to the right of the correlation table respectively. The last row of the table indicates the estimated median intensity percentile of each content column. The percentile metric is used for both content and intensity so that the people are considered to be arranged from 0 to 100 per cent according to their rank on intensity. Each plotted point on the graph corresponds to the midpoint of the content interval and to the intensity percentile of the median case for each content interval.

In practice, the values for the plotted points are determined quite simply. The cumulative percentages are computed for the total frequencies of both content and intensity, cumulating from negative to positive on content and from low to high on intensity. To determine the content value to be plotted we compute the midpoint of the interval of percentiles for each content rank. The intensity value to be plotted is the median intensity percentile for the content rank, as just discussed.

The procedure for studying the intensity function, thus, is very simple. It consists of (1) asking a series of questions with the content of interest; (2) testing these questions to determine whether they all ask the respondents about the same single issue (this is done very simply by means of scalogram boards or any of the other techniques for carrying out a scalogram analysis that have come out of the work of the Research Branch); (3) obtaining for each respondent a content score based upon his rank order on the content scale (if the area is scalable; otherwise there is no point to the further steps); (4) asking a simple intensity question such as, "How strongly do you feel about this?" after each content question in order to obtain an intensity score for each respondent; and (5) plotting content scores by intensity scores to obtain a *U*- or *J*-shaped curve, the lowest point of which serves to divide the population into the desired *objective* and *invariant* "favorable" and "unfavorable" groups.

This technique can best be illustrated by means of the actual examples in the next section.

Illustrations of the Intensity Function

For the most part the following examples will deal with the same attitude areas used to illustrate the scalogram board technique of scale analysis. It will therefore not be necessary to repeat the content scale pictures presented previously. For these pictures and for the complete wording of the questions asked, the reader can refer to Chapter 5. The present section will give only the wording of the content question itself without listing the answer categories. However, the complete wording of the intensity questions and their scalograms will be given. These intensity scalograms are constructed and analyzed in exactly the same manner as the content scalograms and so no additional explanation need be given. The following are the areas of content to be used as illustrations here:

Table	*Universe*
1b.	Attitude toward Officers
2.	Attitude toward the Army
3.	Knowledge of Current Events
4.	Attitude toward Administration of the Army Score Card Plan
5.	Attitude toward Idea of the Army Score Card Plan
6.	Attitude toward the WAC
7.	Attitude toward Postwar Conscription
8.	Satisfaction with One's Army Job

These illustrations will serve to show that a U curve can be determined empirically from content and intensity scores. However a great deal more work must be done on eliminating some of the error from these curves. The present research is only a first approach toward the measurement of the intensity function.

Attitude toward officers. Using the technique outlined above, an intensity analysis was made for the attitude area of "Enlisted Men's Attitudes toward Their Officers." A scalogram of this area showing the questions asked and the scale pattern of responses was given in Chapter 5 (Scalogram 2). Each content question was followed by a question asking about the intensity of feeling. Scalogram 11 shows the intensity questions that were asked after each content question and the scalogram pattern which resulted. The scale analysis and rank order of individuals is determined independently for content and intensity, and consequently there is no reason for the number of scale types to be the same for the content and intensity scales. Since the sensitivity of the intensity curve is a function of the number of content and intensity scale scores that can be correlated, it is advisable in an intensity analysis to keep separate as many answer categories as possible so that the number of scale types is increased.

From Scalogram 11 it can be seen that the responses to the intensity questions form a quasi scale. The meaning of a quasi scale has already been discussed in some detail. Since the reproducibility of responses from scale scores does not apply to quasi scales, we do not compute the coefficient of reproducibility.

Each of the content and intensity questions were dichotomized in the present example. Since there were eleven content questions,

and nine of these were followed by intensity questions, we have twelve content ranks and ten intensity ranks. Each soldier receives two scores: a content score ranging from 0 to 11 depending upon the number of questions upon which he holds "positive" attitudes toward officers, and an intensity score ranging from 0 to 9 depending upon the number of answers about which he feels "strongly." Tabulating intensity score against content score for all soldiers in the entire sample of 2,827 men yields the joint frequency distribution shown in Table 1b.

TABLE 1b

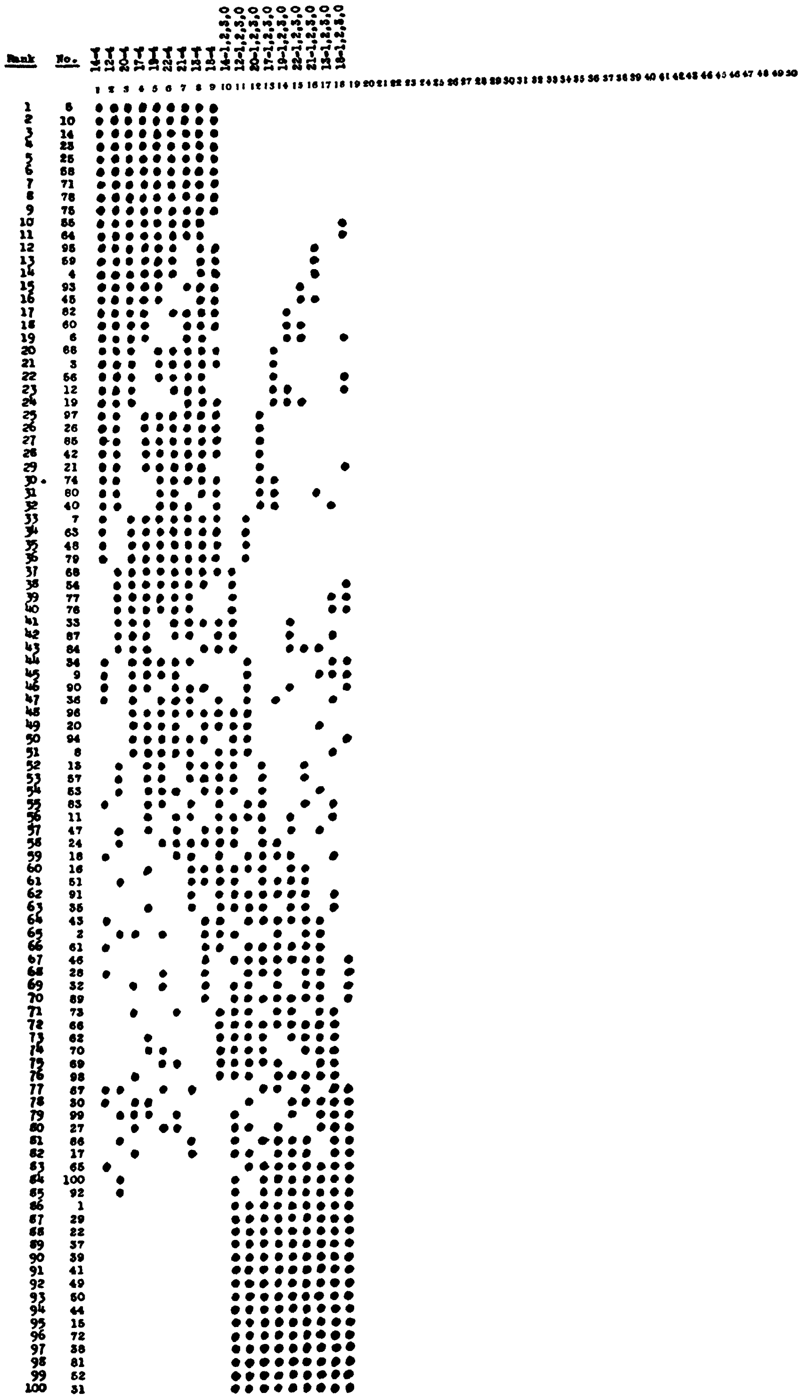
ENLISTED MEN'S ATTITUDES TOWARD THEIR OFFICERS
(Content Scores by Intensity Scores)

Intensity rank	*Content rank* (*Neg.*) *0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	(*Pos.*) *11*	*Total freq.*	*Cum. per cent*
9 (High)	53	76	64	48	35	30	17	21	20	15	5	17	401	100
8	*21*	*35*	40	29	37	20	28	16	15	12	17	*5*	275	86
7	15	27	44	35	32	20	22	17	18	13	6	4	253	76
6	10	16	*43*	35	29	33	30	24	18	12	*12*	6	268	67
5	10	20	30	*25*	36	35	26	17	15	11	10	1	236	58
4	6	10	32	39	*43*	*39*	41	21	23	14	5	1	274	49
3	1	9	31	34	29	34	*29*	*37*	*33*	*11*	5	2	255	40
2	1	4	12	17	47	52	28	32	31	13	6	1	244	31
1	2	—	8	26	32	32	52	37	32	20	4	3	248	22
0 (Low)	2	8	6	33	32	56	62	65	59	40	9	1	373	13
Total frequency	121	205	310	321	352	351	335	287	264	161	79	41	2,827	
Cumulative per cent	4	12	22	34	46	59	71	81	90	96	99	100		
Midpoint of content percentiles	2	8	17	28	40	53	65	76	85	93	97	99		
Median of intensity percentiles	82	78	66	53	48	40	39	33	33	37	58	79		

If the pure intrinsic intensity function were being measured by the present technique, then Table 1b should show intensity as a perfect *U*- or *J*-shaped function of content. Table 1b shows that this is not the case, that there is much error in the technique; but despite the considerable amount of error, the essential shape of the intensity function is apparent. The italicized frequency in each column of Table 1b corresponds to the intensity rank in which the intensity of the median case lies for each content interval.

SCALOGRAM No. 11. INTENSITY OF ATTITUDE TOWARD OFFICERS

Respondent order *Category order*



The last two rows of Table 1b give the percentiles of the midpoints of content and the percentiles of the medians of intensity, which are the values plotted in Figure 1. The method of computing these two points was described in the previous section. For example,[7]

Scalogram No. 11. Intensity of Attitude toward Officers

Questions and Answer Categories

12a. Did your officers give you a good chance to ask questions as to the reason why things were done the way they were?

b. How strongly do you feel about this?

1 ______ Not at all strongly
2 ______ Not so strongly
3 ______ Fairly strongly
4 ______ Very strongly
0 ______ No answer

13a. How many of your officers took a personal interest in their men?

b. Same as 12b.

14a. Do you think that your officers generally did what they could to help you?

b. Same as 12b.

15a. How well do you feel that your officers understood your problems and needs?

16a. Do you feel that your officers recognized your abilities and what you were able to do?

17a. In general, how good would you say your officers were?

b. Same as 12b.

18a. How many of your officers used their rank in ways that seemed unnecessary to you?

b. Same as 12b.

19a. When you did a particularly good job did you usually get recognition or praise for it from your officers?

b. Same as 12b.

20a. How much did you personally like your officers?

b. Same as 12b.

21a How did you feel about the officers that had been selected by the Army?

b. Same as 12b.

22a. How much did you personally respect your officers?

b. Same as 12b.

---

[7] Rounded numbers only are used in this example; for the actual computations, all percentages, including cumulative ones, were carried to three or four significant digits. Of course, there are other equivalent ways of arranging the arithmetic than that presented here.

the midpoint of content percentile for content rank 7 is 76, arrived at as follows:

$$\text{Cum. \% of Content Rank 6} + \frac{\begin{pmatrix}\text{Cum. \% of}\\\text{Content Rank 7}\end{pmatrix} - \begin{pmatrix}\text{Cum. \% of}\\\text{Content Rank 6}\end{pmatrix}}{2}$$

or,

$$71 + \frac{81 - 71}{2} = 76$$

The median of intensity percentile for content rank 7 is 33, arrived at as follows:

$$\text{Cumulative \% of Intensity Rank 2} + \left\{\frac{\dfrac{\begin{pmatrix}\text{Total}\\\text{Freq. of}\\\text{Content}\\\text{Rank 7}\end{pmatrix}}{2} - \begin{pmatrix}\text{Cum. Freq. of}\\\text{Intensity Ranks}\\\text{0, 1, and 2 for}\\\text{Content Rank 7}\end{pmatrix}}{\begin{matrix}\text{Frequency of Intensity Rank 3}\\\text{for Content Rank 7}\end{matrix}}\right\} \left\{\begin{pmatrix}\text{Cum. \% of}\\\text{Intensity}\\\text{Rank 3}\end{pmatrix} - \begin{pmatrix}\text{Cum. \% of}\\\text{Intensity}\\\text{Rank 2}\end{pmatrix}\right\}$$

or,

$$31 + \left(\frac{\frac{287}{2} - 134}{37}\right)(40 - 31)$$

$$= 31 + \left(\frac{10}{37}\right)(9)$$

$$= 33$$

The lowest plotted points fall at 76 and 85 per cent, and the bottom of the curve is fairly flat. The exact location of the zero point in the case of a flat curve is difficult to determine. However, the fact that a large range of "neutrality" exists is in itself an important finding. In particular, *the region in which the zero point lies is invariant.* While it may not be possible to ascertain exactly where the zero point is, we can say that according to this sample of soldiers, 53 per cent of the enlisted men had an unfavorable attitude toward officers; 7 per cent had a favorable attitude toward officers; and the remaining 40 per cent were in between. As more and more questions are added to the scale, the size of the zero range will become more narrow and the location of the zero point can be determined with greater accuracy.

The fact that the curve in Figure 1 is suspended in mid-air and does not reach the bottom percentile of intensity is due to the presence of error in the technique for ascertaining intensity. The col-

umn medians are being plotted; since there is error, the medians are away from where they would be if there were no error. The essential shape of the intensity curve that would be obtained if there were no error seems rather apparent from Figure 1, and it seems safe to assume that the pure curve would actually touch bottom in a region within the zero interval obtained from the observed curve.

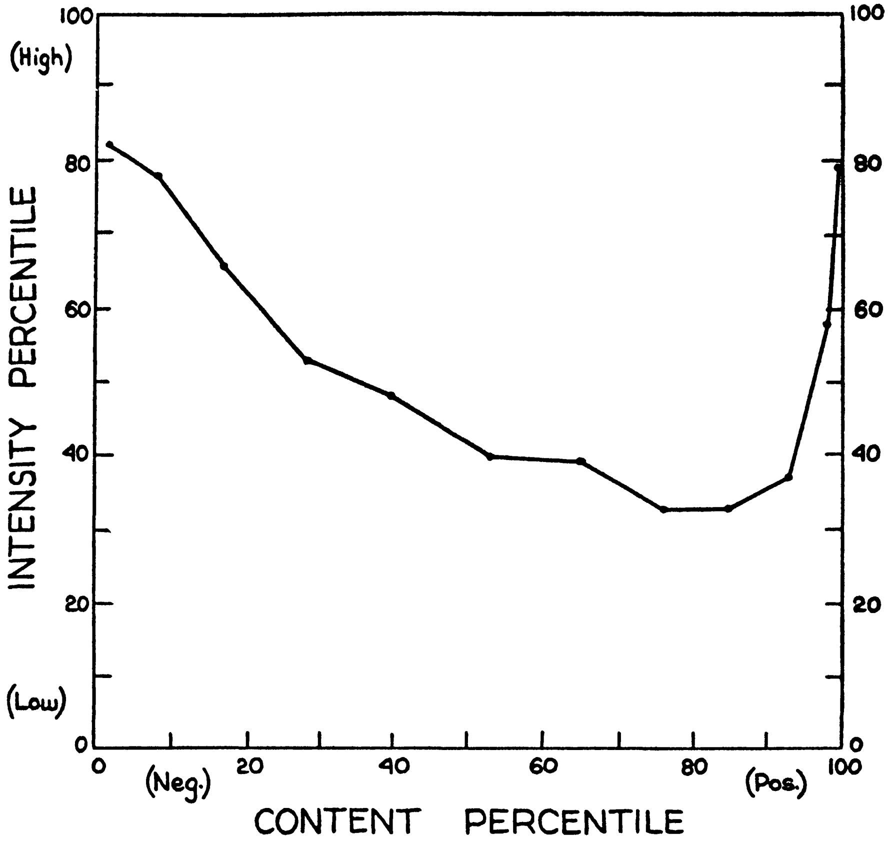
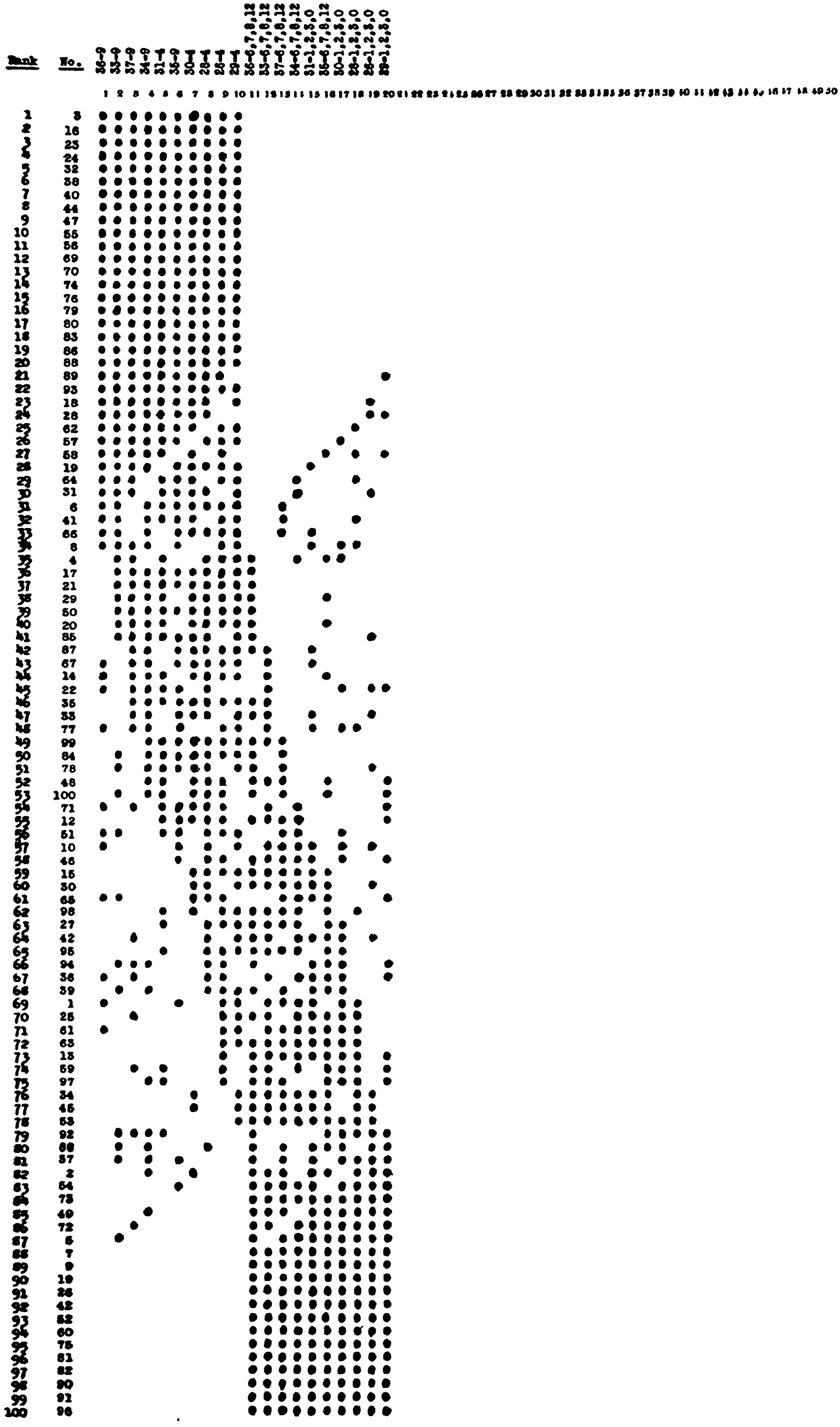


Figure 1. Enlisted men's attitudes toward their officers.

A technique for obtaining a *single point approximation* for the zero point is to use the *median content percentile of the people lowest on intensity*. In the perfect curve, the lowest person on intensity would be at the zero point on content. Since we have error present in practice, we can take an average of the content positions of the people lowest on intensity as an approximation to the ideal zero point. As small a percentage as possible should be used in order to avoid distortion, but the sample number of people included must

SCALOGRAM NO. 12. INTENSITY OF ATTITUDES TOWARD THE ARMY

Respondent order *Category order*



be large enough to be reliable. If we consider only the bottom row in Table 1b (intensity rank 0) the median content percentile of the 373 men therein is, by interpolation, 68 per cent, and we can use this as an estimate of the true zero point. This estimate does fall in the zero interval just previously obtained from the *U* curve.

Scalogram No. 12. Intensity of Attitudes toward the Army

Questions and Answer Categories

26a. All things considered, do you think the Army is run about as efficiently as possible, or do you think it could be run better?
 b. How strongly do you feel about this?
 1 ______ Not at all strongly
 2 ______ Not so strongly
 3 ______ Fairly strongly
 4 ______ Very strongly
 0 ______ No answer

28a. In general do you think the Army has tried its best to see that men get as square a deal as possible?
 b. Same as 26b.

29a. In general, do you feel you yourself have gotten a square deal from the Army?
 b. Same as 26b.

30a. Do you feel that the Army is trying its best to look out for the welfare of enlisted men?
 b. Same as 26b.

31a. In general, how interested do you think the Army is in your welfare?
 b. Same as 26b.

33a In the Army, some jobs are naturally harder and more dangerous than others and the Army has to put men where it thinks they are needed.
 Considering everything, do you think the Army is trying its best to see that, as far as possible, no man gets more than his fair share of the hard and dangerous jobs?
 b. How strongly do you feel about this?
 6 ______ Not at all strongly
 7 ______ Not so strongly
 8 ______ Fairly strongly
 9 ______ Very strongly
 12 ______ No answer

34a. Do you think the Army is trying its best to see that the men who have the hard and dangerous jobs get the special consideration and breaks they deserve?
 b. Same as 33b.

35a. On the whole, do you think the Army gives a man a chance to show what he can do?
 b. Same as 33b.

36a. In general, how well do you think the Army is run?
 b. Same as 33b.

37a. Do you think when you are discharged you will go back to civilian life with a favorable or unfavorable attitude toward the Army?
 b. Same as 33b.

TABLE 2

ENLISTED MEN'S ATTITUDE TOWARD THE ARMY
(Content Scores by Intensity Scores)

Intensity rank	*Content rank* (*Neg.*) *0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	(*Pos.*) *16*	*Total freq.*	*Cum. %*
10 (High)	*46*	26	30	46	58	68	91	73	64	50	37	40	32	28	30	22	*56*	797	100
9	8	*11*	*13*	*20*	25	31	35	57	40	29	19	19	23	20	20	19	16	405	78
8	3	8	11	15	*15*	*29*	44	41	33	38	26	20	22	19	15	*17*	7	363	67
7	1	7	10	11	22	17	*38*	35	35	34	25	20	11	12	*14*	11	4	307	57
6	—	1	7	9	7	19	31	*42*	*36*	23	26	16	*13*	*14*	10	8	5	267	48
5	—	1	6	7	13	20	27	46	37	*27*	*26*	*23*	12	11	5	5	2	268	41
4	1	—	—	10	9	8	16	37	29	23	18	27	19	8	7	1	1	214	33
3	—	—	1	3	10	9	24	26	25	28	29	26	14	12	10	4	—	221	27
2	—	—	—	1	3	10	21	40	28	20	33	20	15	15	7	1	—	214	21
1	—	—	—	2	1	5	21	24	30	38	29	25	13	12	7	2	2	211	15
0 (Low)	—	—	—	5	7	7	24	65	52	47	43	31	25	25	6	7	3	347	10
Total frequency	59	54	78	129	170	223	372	486	409	357	311	267	199	176	131	97	96	3,614	
Cumulative per cent	2	3	5	9	14	20	30	43	55	65	73	81	86	91	95	97	100		
Midpoint of content percentiles	1	2	4	7	11	17	25	37	49	60	69	77	83	89	93	96	99		
Median of intensity percentiles	86	77	70	68	65	62	53	42	42	40	34	35	42	43	56	62	81		

Attitudes toward the Army. A second illustration of the intensity function concerns the area of enlisted men's attitudes toward the Army. The scale analysis of the content questions in this area was presented in detail in Chapter 4, "The Scalogram Board Technique for Scale Analysis." The scale pattern of the intensity questions which followed each of the content questions can be seen from Scalogram 12.

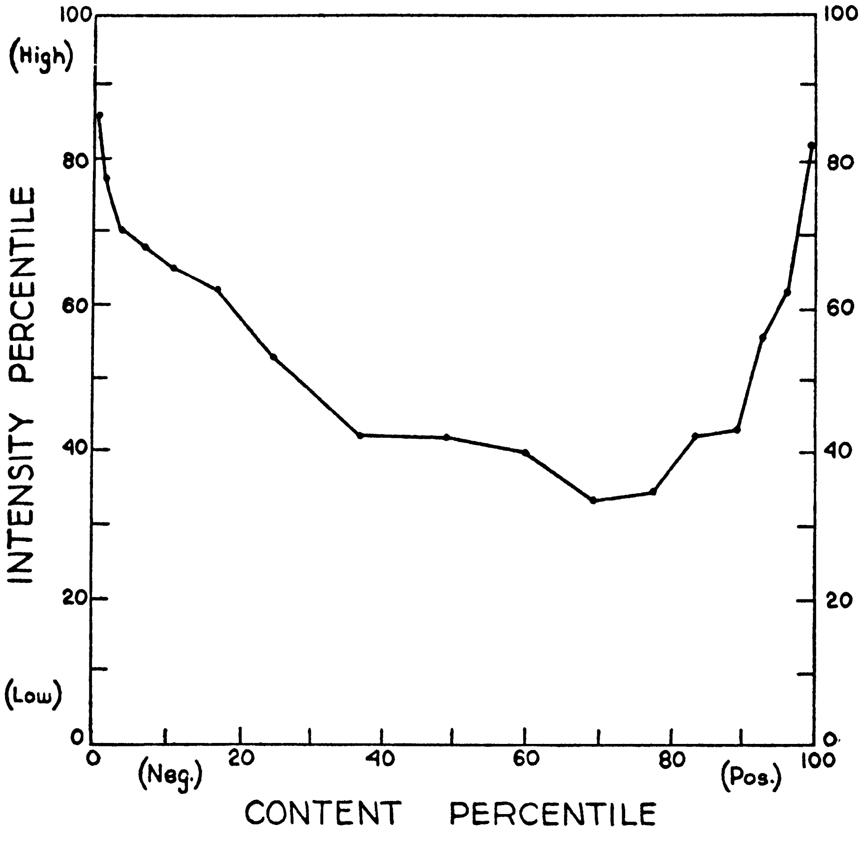
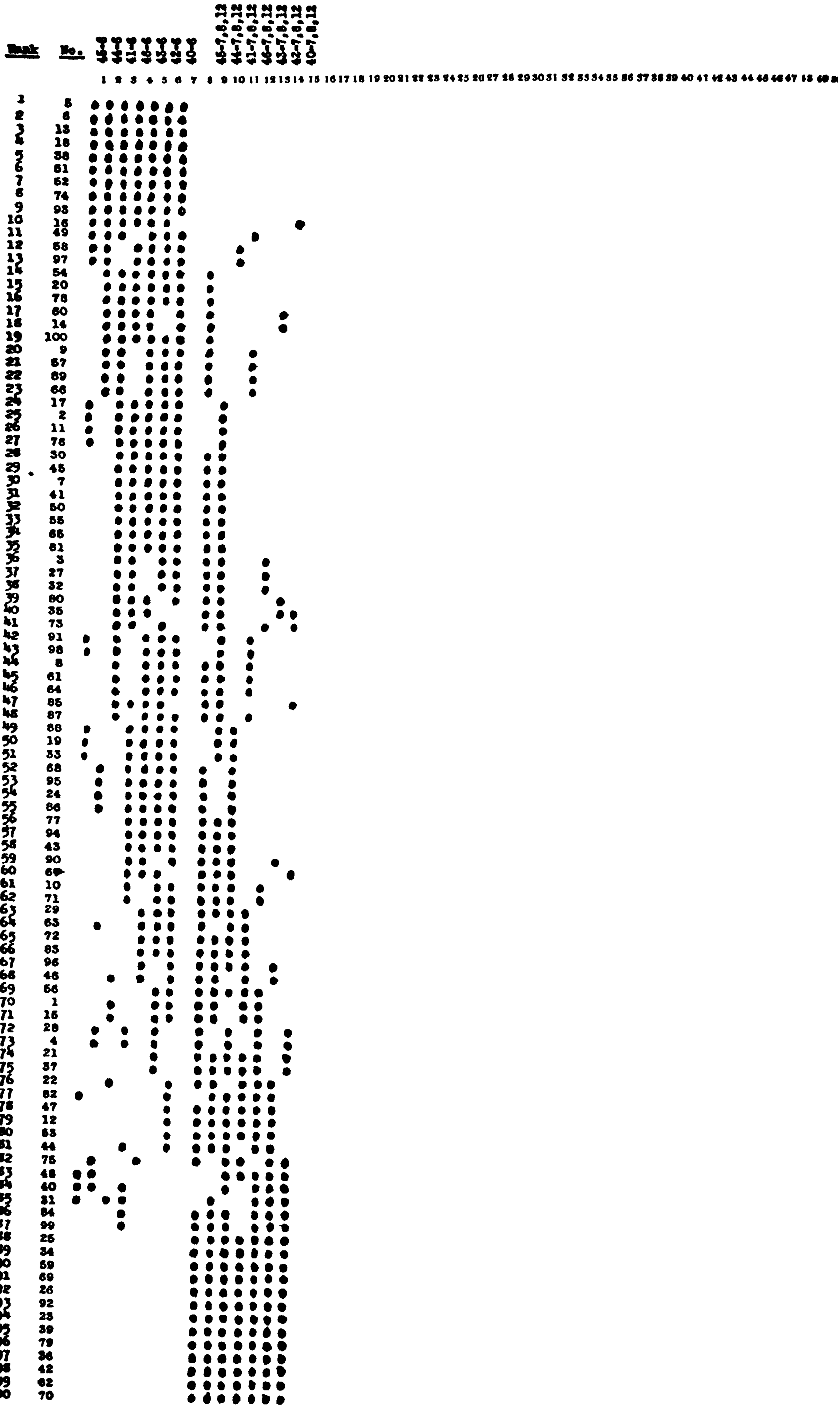


Figure 2. Enlisted men's attitude toward the Army.

Once again we find the familiar quasi-scale pattern of intensity. Each of the ten intensity questions was dichotomized, producing eleven intensity ranks. Several of the content questions were left as trichotomies making it possible to differentiate between seventeen meaningful content types. The tabulation of intensity scores against content scores is given in Table 2. Figure 2 shows the intensity curve resulting from the plotting of median intensities.

SCALOGRAM NO. 13. INTENSITY OF KNOWLEDGE OF CURRENT EVENTS

Respondent order *Category order*



As in the case of attitudes toward officers, this curve is definitely skewed. The lowest obtained point falls at 69 per cent, but once again the zero range is fairly broad. We can safely say that 37 per cent are negative as compared to 11 per cent positive, while 52 per cent are in between.

Knowledge of current events. Scale analysis is useful not only for the investigation of attitudes, but for any universe of content which is observed as qualitative data. Information tests are important sets of qualitative data which require scale analysis if they are to be scored meaningfully.

How are the intensity function and zero point related to information tests? Our customary verbalizations about information tests do not seem to consider "positive" and "negative" knowledge. Knowledge seems usually to be thought of as nonnegative. That this may often be justified is illustrated by the following example.

The universe of content used was knowledge of facts associated with the war. The scalogram for this area is given in Chapter 5, Scalogram 6. Intensity was asked by following each content question with a question inquiring as to degree of certainty of the answer. Scalogram 13 shows the scale pattern obtained by the responses to these intensity questions.

Scalogram No. 13. Intensity of Knowledge of Current Events

Questions and Answer Categories

40a. Who is the Chief of Staff of the U.S. Army?
  b. How sure are you of your answer?
    6 ______ Very sure
    7 ______ Fairly sure
    8 ______ Not sure
    12 ______ No answer

41a. Which of the following countries is at war against the Axis?
  b. Same as 40b.

42a. Where was the first major beach head established by the Allies in their invasion of Europe?
  b. Same as 40b.

43a. Who is the recently appointed Secretary of State?
  b. Same as 40b.

44a. The man who said we would have "peace in our time" was ________?
  b. Same as 40b.

45a. The "Co-Prosperity Sphere" was a scheme of ________?
  b. Same as 40b.

46a. "Fascism" began in ________?
  b. Same as 40b.

The tabulation of content scores by intensity scores is given in Table 3, while the intensity curve is shown in Figure 3. This curve has an interesting shape. It consists of but one arm of the *U*, and every content score may be said to be on the positive side of zero. Certainty increases as knowledge increases, while the less one knows, the less certain one is of one's answers. This is in sharp contrast to attitudes where *both* positive and negative answers can be held with strong intensity. It would seem that if a person gives the wrong answer to an apparently factual question and yet is firmly convinced that his answer is right, there is a strong possibility that the question is not tapping an informational area but an attitudinal area.

Additional examples. The shape of the intensity curve can vary considerably from area to area. This will be seen in the remainder of the examples to be given. However since the reader by now should be well acquainted with the scale pictures for the intensity function, we will not present the separate scalograms for the remaining examples. Suffice it to say that in each case content formed a scale and intensity proved to be a quasi scale. The tabulation of intensity scores by content scores will be presented, together with the intensity curves.

Attitude toward administration of Army score card plan. An interesting example of intensity analysis is afforded by a study of soldiers' opinions of the Army's demobilization score card plan. Two areas were to be studied. One was opinion of the *idea* of the score card as a demobilization plan and the other was opinion of *how the plan was being carried out in practice.* Let us consider this second area first. Six questions were asked as a sample of the universe of content. As in the previous examples, each question was actually in two parts; the first part concerned the content, and the second part elicited intensity of feeling. One of the questions was, for example:

(a) In general, do you think the Army is trying its best to carry out the Army score card plan as it should be carried out?
______ Yes, it is trying its best
______ It is trying some, but not hard enough
______ It is hardly trying at all

(b) How strongly do you feel about this?
______ Not at all strongly
______ Not so strongly
______ Fairly strongly
______ Very strongly

TABLE 3

KNOWLEDGE OF CURRENT EVENTS
(Content Scores by Intensity Scores)

Intensity rank	*Content rank* (*Neg.*) *0*	*1*	*2*	*3*	*4*	*5*	*6*	(*Pos.*) *7*	*Total freq.*	*Cum. %*
7 (High)	—	—	—	4•	2	6	5	15	32	100
6	—	—	1	2	2	5	12	*7*	29	89
5	—	1	2	4	4	15	*23*	9	58	79
4	1	—	1	4	10	*10*	8	2	36	60
3	—	—	6	*8*	*9*	7	3	4	37	47
2	1	4	*5*	7	8	7	3	1	36	35
1	1	*8*	4	3	5	1	—	1	23	23
0 (Low)	*15*	8	6	7	5	3	—	—	44	15
Total frequency	18	21	25	39	45	54	54	39	295	
Cumulative per cent	6	13	22	35	50	68	87	100		
Midpoint of content percentiles	3	10	17	28	43	59	78	93		
Median of intensity percentiles	9	17	29	39	41	58	71	83		

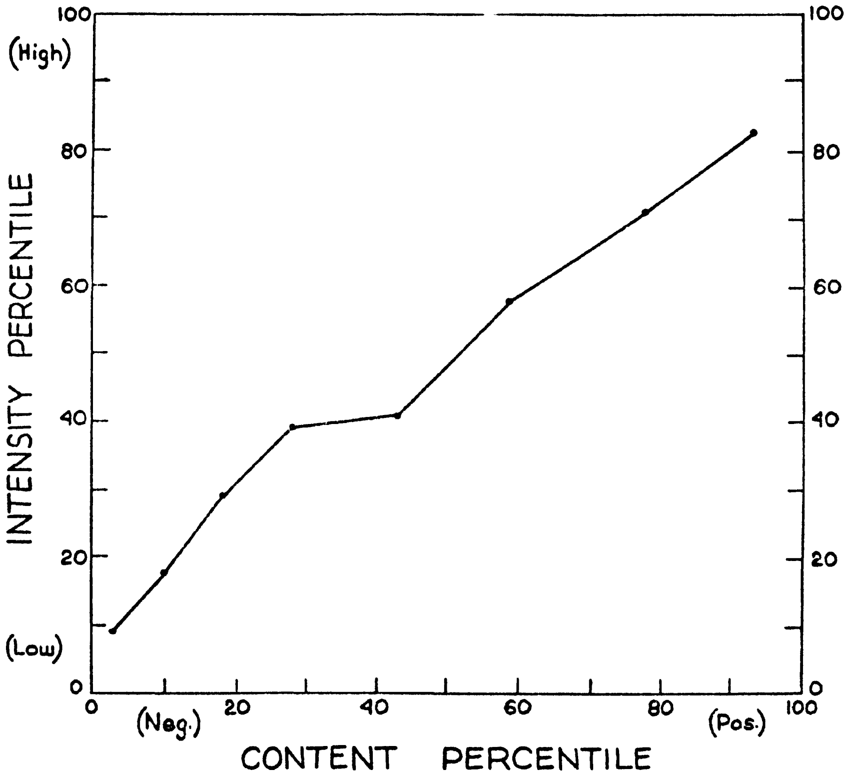


Figure 3. Knowledge of current events.

Each part *a* of the remaining five questions asked a different opinion question about how the plan was being carried out; each part *b* was identical with the above for all the questions.

The joint distribution of content and intensity ranks for the attitude of soldiers on how the score card plan was being carried out is given in Table 4, and the approximate intensity function is plotted in Figure 4. The expected *U* curve of intensity on content results with the low point of the curve, the zero point, falling between percentiles 28 and 67 of the content scale. This indicates a split among the soldiers in their attitude toward the *way in which the score card plan was being administered,* with about half the soldiers having a favorable attitude and half having an unfavorable attitude. In general, the curve has a rather flat bottom indicating a wide zero range or area of indifference. The shape of this curve is quite different from the positively skewed curves of attitude toward officers and attitude toward the Army.

Attitude toward idea of Army score card plan. The other aspect of the problem, the soldier's attitude toward the *idea* of the score card plan, produced an altogether differently shaped intensity curve. Eight questions, containing the content in part *a* and the intensity in part *b*, were asked as a sample of all questions that could have been asked in this area.

The joint distribution of content and intensity ranks for soldiers' attitudes toward the *idea* of the score card plan is given in Table 5, and the approximate intensity function is plotted in Figure 5. The zero point seems to be indefinitely far to the left. This indicates that there was little or no unfavorable opinion about the idea of the score card plan—practically everybody was favorable. Such a conclusion could not be ascertained by looking at the marginal frequencies of the individual questions asked in this area. For example, one of the questions was: "In general, what do you think of the Army score card plan (the point system)?" The responses were as follows:

It is very good	23%
It is fairly good	49
It is not so good	16
It is not good at all	7
Undecided	5
	100%

It is important to notice that the technique of single point estimation of the zero point by using the content median of only the

SOLDIER OPINION OF THE ADMINISTRATION OF THE SCORE CARD
(Content Scores by Intensity Scores)

Intensity rank	Content rank (Neg.) 0	1	2	3	4	5	(Pos.) 6	Total freq.	Cum. per cent
9 (High)	65	50	42	22	30	33	81	323	100
8	*44*	66	70	40	45	52	*57*	374	90
7	39	55	84	60	59	63	44	404	79
6	16	*67*	98	85	84	85	25	460	66
5	17	42	*93*	*114*	*112*	*91*	28	497	52
4	12	34	130	128	157	157	11	629	37
3	10	21	64	82	70	19	1	267	17
2	—	8	46	46	28	9	1	138	9
1	4	8	19	28	13	6	1	79	5
0 (Low)	4	18	23	15	7	4	—	71	2
Total frequency	211	369	669	620	605	519	249	3,242	
Cumulative per cent	7	18	39	58	76	92	100		
Midpoint of content percentiles	3	12	28	48	67	84	96		
Median of intensity percentiles	79	63	45	38	40	47	81		

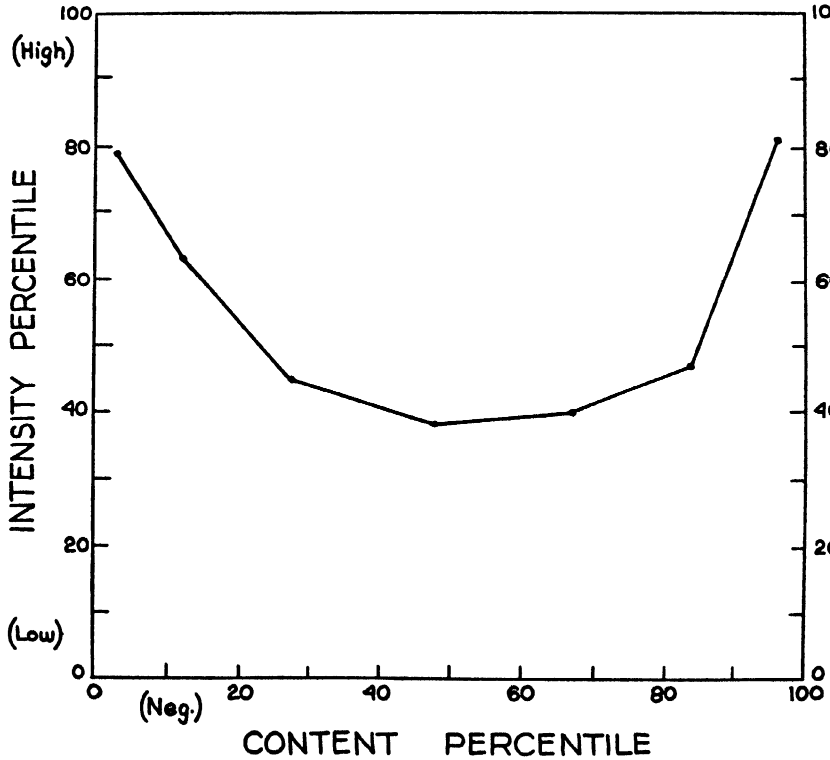


Figure 4. Soldier opinion of the administration of the score card.

TABLE 5

SOLDIER OPINION OF THE IDEA OF THE SCORE CARD
(Content Scores by Intensity Scores)

Intensity rank	*Content rank* (*Neg.*) *0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	(*Pos.*) *8*	*Total freq.*	*Cum. %*
8 (High)	26	58	75	26	17	30	46	54	*104*	436	100
7	15	62	60	27	16	28	34	*45*	12	299	87
6	18	46	73	32	19	30	*34*	19	8	279	77
5	13	54	102	32	36	*31*	10	10	4	292	69
4	19	55	113	67	*38*	28	18	5	—	343	60
3	20	*91*	*144*	*84*	42	23	11	3	2	420	49
2	*15*	80	157	66	29	22	7	3	3	382	36
1	32	119	177	70	27	8	6	2	—	441	24
0 (Low)	69	134	80	37	15	10	3	2	—	350	11
Total frequency	227	699	981	441	239	210	169	143	133	3,242	
Cumulative per cent	7	29	59	72	80	86	91	96	100		
Midpoint of content percentiles	4	18	44	66	76	83	89	94	98		
Median of intensity percentiles	34	39	43	44	51	64	76	83	91		

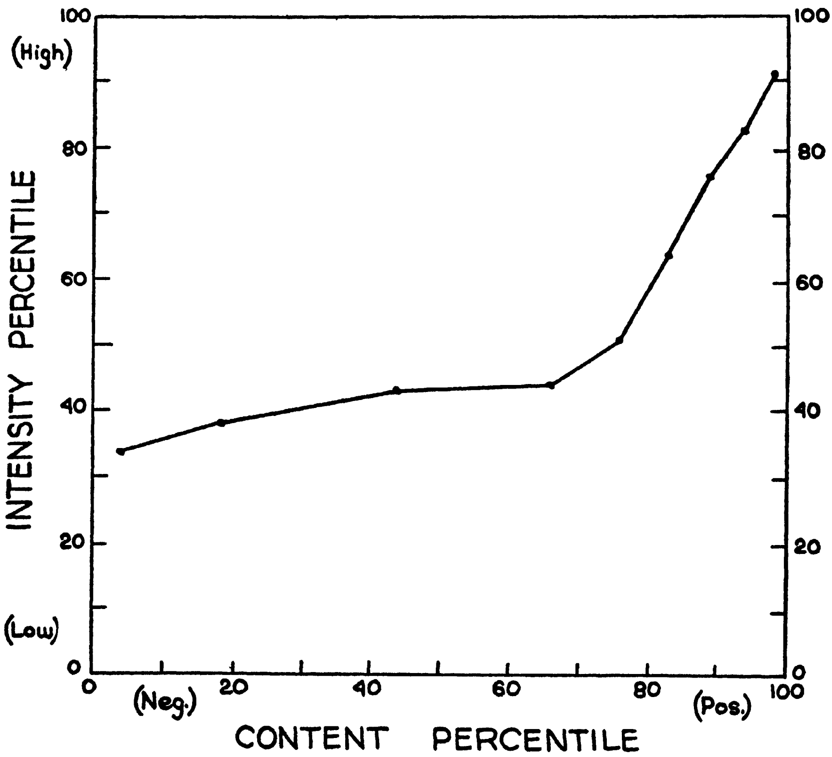


Figure 5. Soldier opinion of the idea of the score card.

lowest intensity group breaks down in this instance. Using the technique of lowest intensity is not safe if the intensity function is too asymmetric.

Attitude toward the WAC. An example of a very sharp intensity curve is afforded by the universe of soldiers' attitudes toward the WAC. The content scalogram and the sample of questions used are given in Chapter 5, Scalogram 3. The tabulation of intensity scores against content scores is given in Table 6, while the intensity curve is presented in Figure 6.

From this curve we see that there is a very definite zero point at 75 per cent. The sharpness of this curve would indicate that there are few soldiers who are "neutral" in their attitudes toward the WAC, and that the ratio of approval to disapproval is approximately 1 to 3. Very sharp *U* curves indicate a clear distinction between being positive and being negative.

Attitude toward postwar conscription. Another example of a fairly sharp *U* curve is found in the area of soldiers' attitudes toward postwar conscription. The content of this area can be seen from Scalogram 4, Chapter 5. Table 7 and Figure 7 give the tabulation and graph of content scores by intensity scores.

The low point of this intensity curve is in the negative region of the content continuum, showing more soldiers in favor of postwar universal military training than opposed to such training.

Satisfaction with one's job. A final example of an extremely flat-bottomed intensity curve is that of soldiers' attitudes toward their jobs. The content scalogram and questions asked in this area are given in Chapter 5, Scalogram 1. The tabulation of content by intensity scores is given in Table 8, while the intensity curve is drawn in Figure 8.

There is a wide range of indifference running from the 34th percentile to the 77th percentile. One conclusion from this curve would be that soldiers are not sharply divided in regard to satisfaction with their jobs. There is an appreciably large group that can be labeled as neither satisfied nor dissatisfied.

Relativity of the Intensity Function

As in the case of content scale analysis, the intensity function is relative to the population and the time of the study. This is to be expected since the intensity function is as intrinsic a part of the attitude universe as the content scale. For a detailed discussion of the relativity of an attitude universe, the reader is referred to Chapter 5.

TABLE 6

ENLISTED MEN'S ATTITUDE TOWARD THE WAC
(Content Scores by Intensity Scores)

Intensity rank	*Content rank* (*Neg.*) *0*	*1*	*2*	*3*	*4*	*5*	(*Pos.*) *6*	*Total freq.*	*Cum. %*
6 (High)	*92*	78	47	21	21	13	20	292	100
5	37	50	34	21	21	6	9	178	83
4	17	*50*	46	22	28	11	10	184	73
3	26	27	*65*	39	36	13	*15*	221	62
2	4	22	65	*60*	60	*27*	18	256	49
1	4	20	48	45	*123*	34	10	284	35
0 (Low)	3	12	61	59	146	30	4	315	18
Total frequency	183	259	366	267	435	134	86	1,730	
Cumulative per cent	11	26	47	62	87	95	100		
Midpoint of content percentiles	5	18	36	54	75	91	98		
Median of intensity percentiles	83	73	51	42	28	36	59		

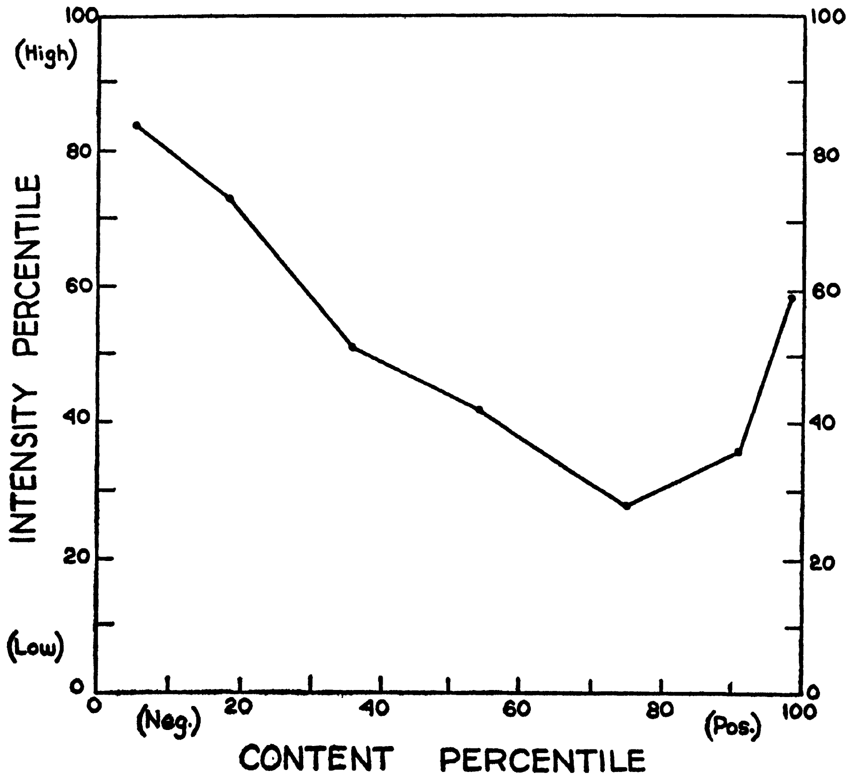


Figure 6. Enlisted men's attitude toward the WAC.

TABLE 7

Attitude toward Postwar Conscription
(Content Scores by Intensity Scores)

Intensity rank	*Content rank* (*Neg.*) *0*	*1*	*2*	*3*	*4*	*5*	(*Pos.*) *6*	*Total freq.*	*Cum. per cent*
5 (High)	60	42	23	19	34	55	*174*	407	100
4	*42*	39	17	17	37	58	64	274	77
3	24	39	28	35	54	*53*	26	259	61
2	20	*47*	30	46	*60*	37	12	252	47
1	4	57	*51*	*52*	61	22	5	252	33
0 (Low)	12	60	69	121	49	9	2	322	18
Total frequency	162	284	218	290	295	234	283	1,766	
Cumulative per cent	9	25	38	54	71	84	100		
Midpoint of content percentiles	5	17	31	46	62	77	92		
Median of intensity percentiles	69	40	29	25	41	60	81		

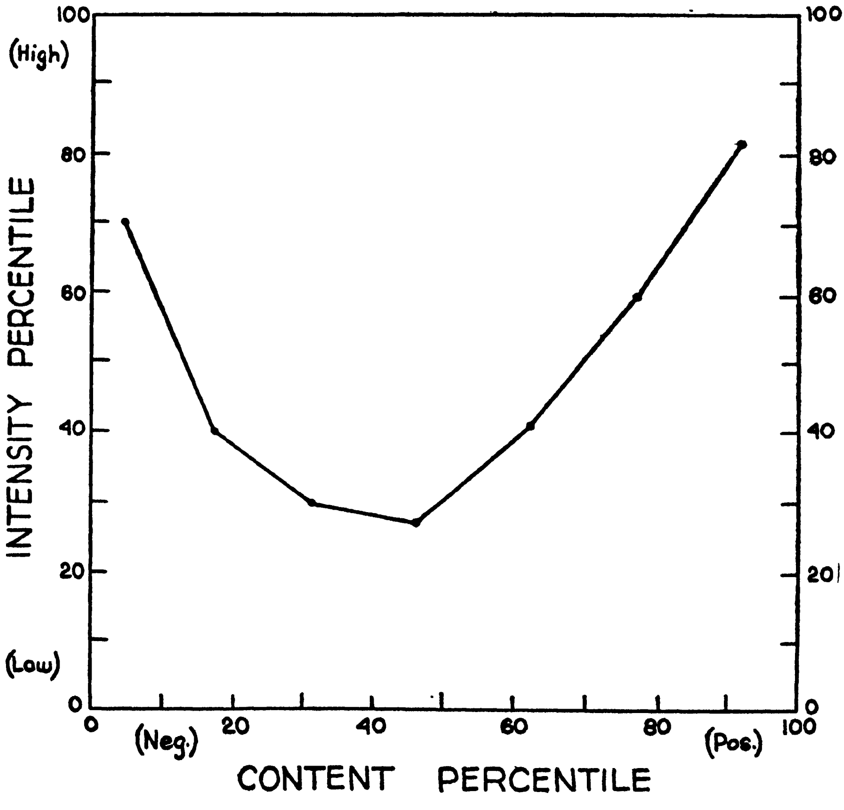


Figure 7. Attitude toward postwar conscription.

TABLE 8

Satisfaction with One's Army Job
(Content Scores by Intensity Scores)

Intensity rank	*Content rank* (*Neg.*) *0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10* (*Pos.*)	*11*	*Total freq.*	*Cum. %*
8 (High)	*23*	46	27	33	22	19	24	20	9	25	23	*24*	295	100
7	7	*24*	31	26	33	31	22	13	20	21	*15*	5	248	84
6	1	7	*29*	17	30	24	35	17	23	*15*	11	—	209	70
5	6	14	14	*29*	20	34	27	13	*17*	9	10	—	193	58
4	2	3	15	17	*32*	*35*	*36*	19	*20*	10	1	—	190	47
3	—	1	17	19	22	*27*	33	*18*	7	11	1	—	156	37
2	1	4	9	19	25	34	31	25	14	1	4	1	168	28
1	—	2	2	12	35	39	38	23	13	8	—	1	173	19
0 (Low)	—	3	7	12	29	43	33	21	15	3	1	—	167	9
Total frequency	40	104	151	184	248	286	279	169	138	103	66	31	1,799	
Cumulative per cent	2	8	16	27	40	56	72	81	89	95	98	100		
Midpoint of content percentiles	1	5	12	22	34	48	64	77	85	92	96	99		
Median of intensity percentiles	86	80	63	52	41	37	38	36	47	66	74	89		

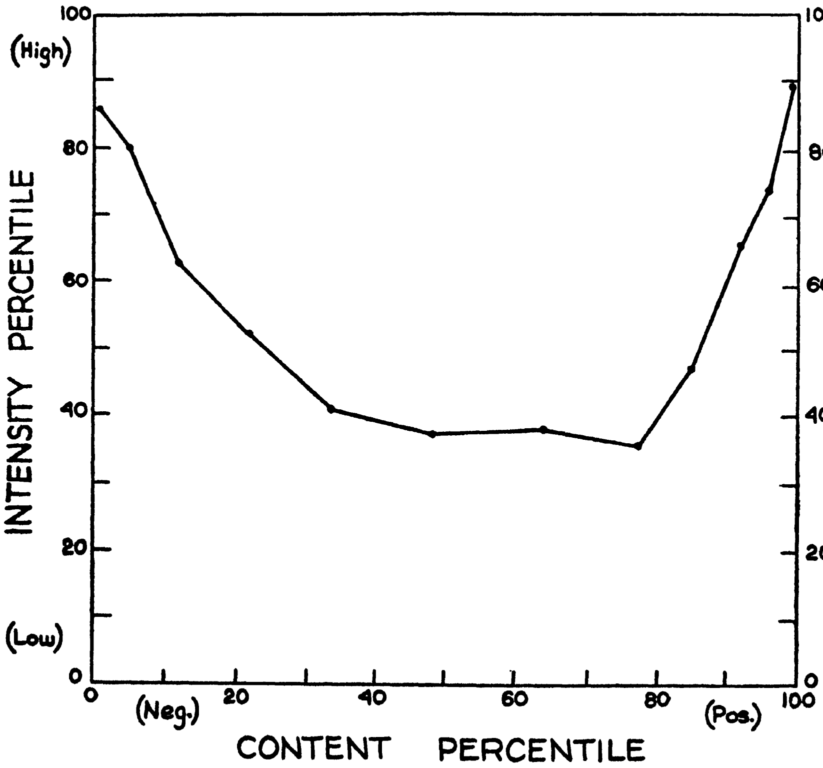


Figure 8. Satisfaction with one's Army job.

In regard to subpopulations of the total population, the intensity function can serve the useful purpose of permitting comparisons between groups as to absolute percentages of various groups that are positive or negative. For example, with the aid of an intensity analysis one can say not only that privates are relatively less satisfied with their jobs than noncoms, but also that there are more privates than noncoms who are dissatisfied. This type of absolute comparison adds important information to our understanding of how groups differ.

The comparison of percentages of a population who are positive and negative can also be made between different attitude or opinion areas. Thus, for example, with an intensity analysis one could compare the number of men who have favorable attitudes toward the British with the number of men who have favorable attitudes toward the Russians.

Several examples of an intensity analysis for subgroups of a population are given below. From Figures 9 and 10, based on Tables 9 and 10, it is possible to conclude that more noncoms than privates have negative attitudes toward the WAC. Both the zero range and the zero point of the noncoms are more to the right than those of the privates.

Figures 11 and 12, based on Tables 11 and 12, compare privates and noncoms in regard to satisfaction with their job. It is quite clear from these curves that more privates than noncoms are dissatisfied with their jobs. While one could conclude from a comparison of content scale scores alone that privates are less satisfied than noncoms, the present comparison permits a direct statement of the absolute percentages dissatisfied in both groups. A striking comparison is given by the intensity function computed separately for men who have completed high school as civilians but who are only privates in rank. The zero point score falls at 89 per cent, showing a high degree of dissatisfaction. These results are shown in Table 13 and Figure 13.

An interesting illustration of the ability of the intensity function to discriminate between groups is shown in Figures 14 and 15, based on Tables 14 and 15. The group of men who do *not* believe that the Army Score Card Plan would be carried out in practice the way it was supposed to be is seen to contain many more men with negative attitudes toward the Army in general, than the group of men who believe that the Army Score Card Plan would be carried out as promised.

TABLE 9: ATTITUDE TOWARD THE WAC—PRIVATES
(Content Scores by Intensity Scores)

Intensity rank	*Content rank* (*Neg.*) *0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	(*Pos.*) *8*	*Total freq.*	*Cum. %*
15 (High)	*5*	2	—	—	—	—	—	1	1	9	100
14	—	6	4	—	—	1	—	1	—	12	98
13	1	4	8	2	1	1	3	1	—	21	96
12	1	*5*	10	7	3	2	3	4	—	35	91
11	—	2	*10*	2	4	4	4	2	—	28	84
10	—	4	11	5	8	7	3	6	*1*	45	78
9	—	—	10	13	7	3	3	*8*	—	44	68
8	—	1	6	*13*	10	8	5	6	—	49	59
7	—	—	1	10	*11*	12	*6*	4	—	44	49
6	—	1	2	8	8	16	8	3	1	47	39
5	—	1	—	9	6	*24*	6	1	—	47	30
4	—	—	—	4	5	30	6	—	—	45	20
3	—	—	—	3	4	17	1	—	—	25	10
2	—	—	—	1	3	7	3	—	—	14	5
1	—	—	—	—	2	3	1	—	—	6	2
0 (Low)	—	—	—	—	—	2	1	—	—	3	1
Total frequency	7	26	62	77	72	137	53	37	3	474	
Cumulative per cent	1	7	20	36	51	80	92	99	100		
Midpoint of content percentiles	1	4	14	28	44	66	86	95	100		
Median of intensity percentiles	99	90	78	52	46	24	40	64	73		

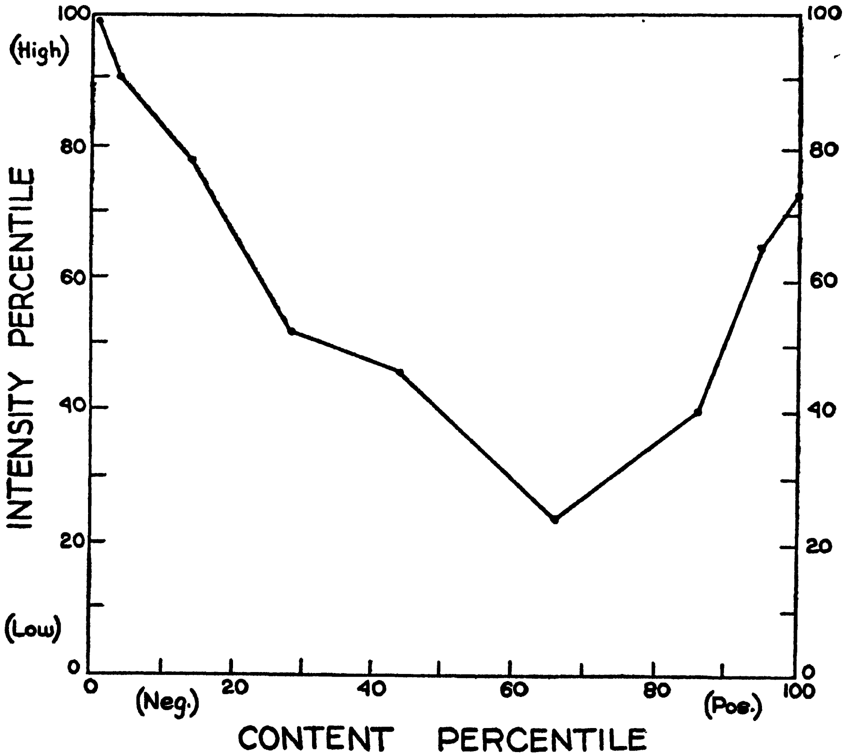


Figure 9. Attitude toward the WAC—privates

TABLE 10: ATTITUDE TOWARD THE WAC—NONCOMS
(Content Scores by Intensity Scores)

Intensity rank	*Content rank* (*Neg.*) *0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	(*Pos.*) *8*	*Total freq.*	*Cum. %*
15 (High)	*8*	3	—	—	—	—	—	—	3	14	100
14	3	10	4	—	—	1	—	—	1	19	97
13	—	*5*	16	1	—	—	—	1	*1*	24	92
12	1	*7*	9	6	3	1	1	2	1	31	87
11	—	4	*15*	15	4	2	—	—	1	41	80
10	—	2	5	13	5	4	2	2	1	34	70
9	—	2	7	18	5	2	1	*4*	—	39	62
8	—	2	6	*11*	11	8	6	—	—	44	53
7	—	—	—	8	*9*	7	2	3	1	30	43
6	—	—	2	16	7	19	*4*	—	—	48	36
5	—	—	2	10	6	*19*	3	—	—	40	25
4	—	—	1	7	6	22	5	—	—	41	16
3	—	—	—	2	3	5	2	—	—	12	7
2	—	1	1	1	4	7	1	—	—	15	4
1	—	—	—	—	—	—	—	—	—	—	—
0 (Low)	—	—	1	—	—	1	—	—	—	2	—
Total frequency	12	36	69	108	63	98	27	12	9	434	
Cumulative per cent	3	11	27	52	66	89	95	98	100		
Midpoint of content percentiles	1	7	19	39	59	78	92	96	99		
Median of intensity percentiles	98	87	76	53	41	23	32	60	90		

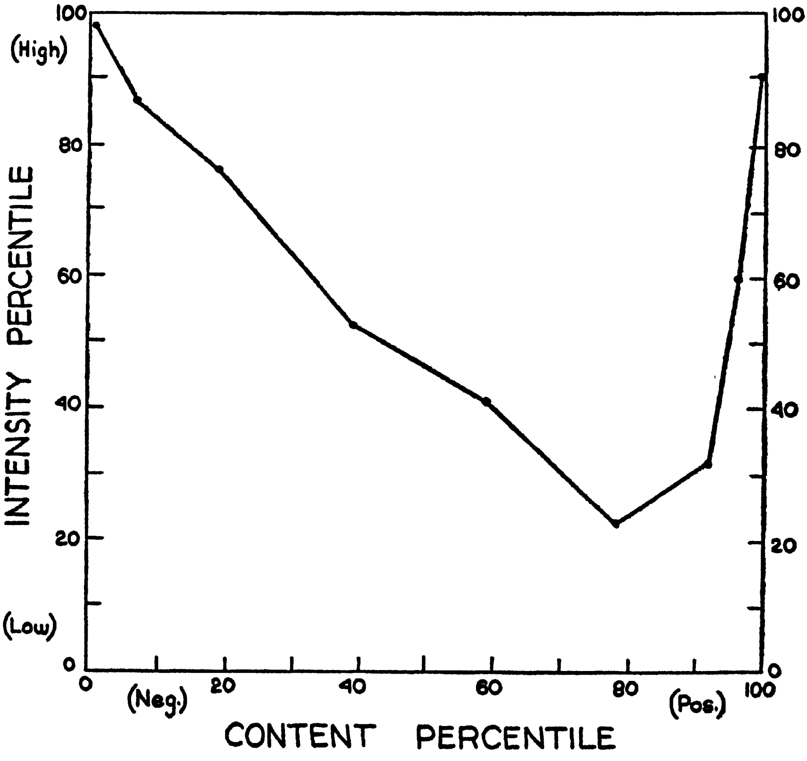


Figure 10. Attitude toward the WAC—noncoms.

TABLE 11

SATISFACTION WITH ONE'S JOB—PRIVATES
(Content Scores by Intensity Scores)

Intensity rank	*Content rank* (*Neg.*) *0–1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	(*Pos.*) *9–12*	*Total freq.*	*Cum. %*
8 (High)	39	10	13	11	10	13	6	3	19	124	100
7	*21*	12	11	23	20	9	4	9	*9*	118	86
6	6	16	13	18	15	25	8	8	7	116	72
5	10	*8*	*17*	9	20	18	7	*4*	5	98	59
4	2	6	8	*19*	*21*	*18*	6	6	4	90	47
3	1	13	7	13	15	18	*8*	2	3	80	37
2	1	8	9	13	22	16	12	3	—	84	28
1	1	1	5	20	20	18	12	5	2	84	18
0 (Low)	3	3	6	10	22	12	7	6	2	71	8
Total frequency	84	77	89	136	165	147	70	46	51	865	
Cumulative per cent	10	19	29	45	64	81	89	94	100		
Midpoint of content percentiles	5	14	24	37	54	72	85	91	97		
Median of intensity percentiles	84	58	54	43	39	42	32	50	76		

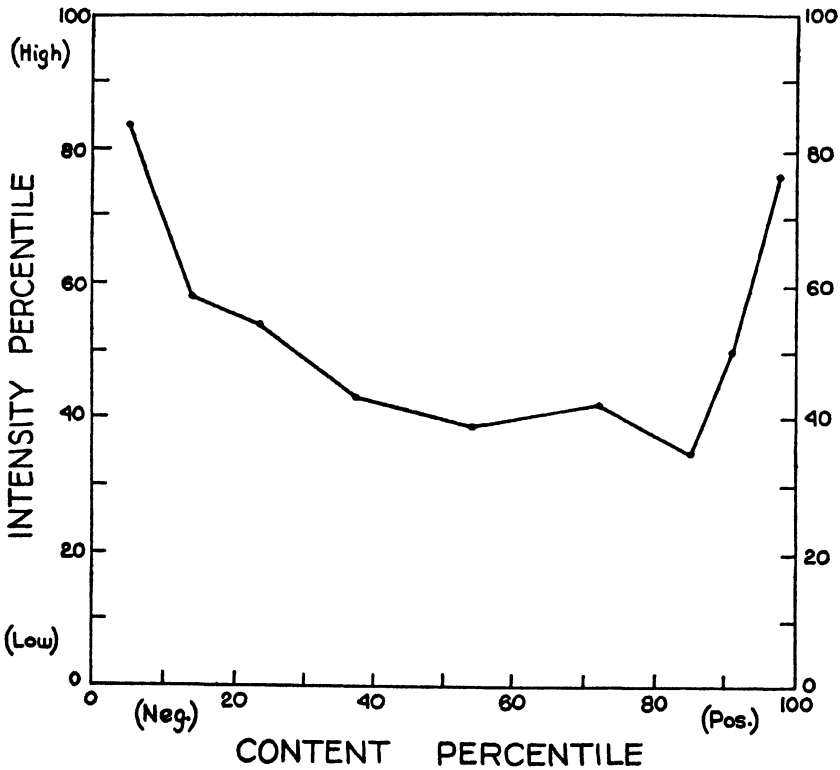


Figure 11. Satisfaction with one's job—privates.

TABLE 12

Satisfaction with One's Job—Noncoms
(Content Scores by Intensity Scores)

Intensity rank	*Content rank* (*Neg.*) *0–1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	(*Pos.*) *9–12*	*Total freq.*	*Cum. %*
8 (High)	*30*	17	20	11	9	11	14	6	53	171	100
7	*10*	19	15	10	11	13	9	11	*32*	130	82
6	2	*13*	4	12	9	10	9	15	19	93	68
5	10	6	*12*	11	14	9	6	13	14	95	58
4	3	9	9	*13*	14	18	*13*	*14*	7	100	48
3	—	4	12	9	*12*	*15*	10	5	9	76	37
2	4	1	10	12	12	15	13	11	6	84	29
1	1	1	7	15	19	20	11	8	7	89	20
0 (Low)	—	4	6	19	21	21	14	9	2	96	10
Total frequency	60	74	95	112	121	132	99	92	149	934	
Cumulative per cent	6	14	25	37	49	64	74	84	100		
Midpoint of content percentiles	3	10	19	31	43	57	69	79	92		
Median of intensity percentiles	82	67	51	38	35	34	38	47	72		

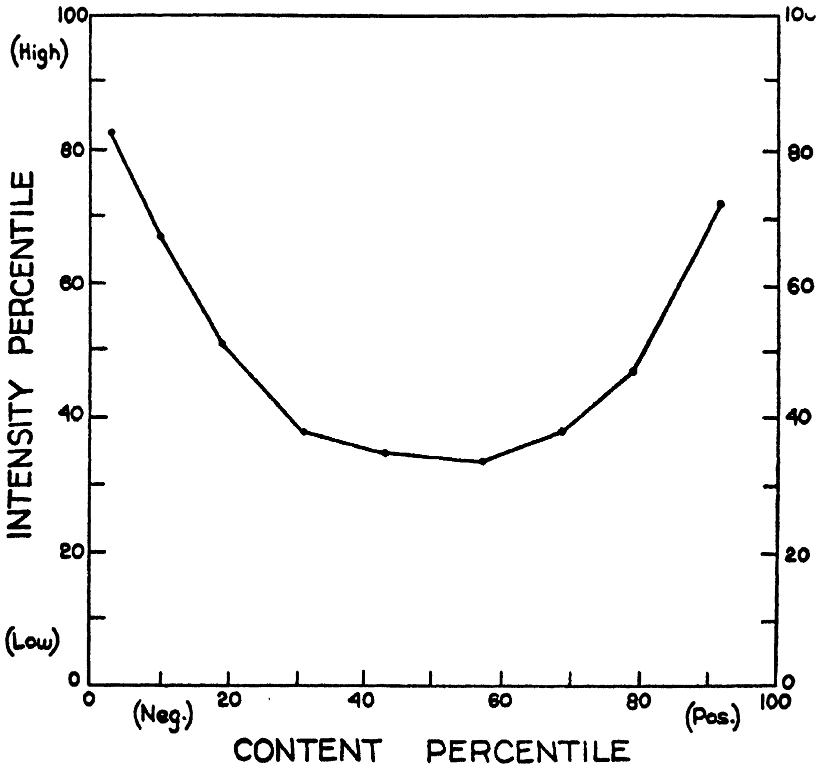


Figure 12. Satisfaction with one's job—noncoms.

TABLE 13

SATISFACTION WITH ONE'S JOB—EDUCATED PRIVATES
(Content Scores by Intensity Scores)

Intensity rank	*Content rank* (*Neg.*) *0–1*	*2*	*3*	*4*	*5*	*6*	*7–8*	(*Pos.*) *9–12*	*Total freq.*	*Cum. %*
8 (High)	12	6	3	3	3	—	2	3	32	100
7	7	3	3	6	4	3	4	3	33	87
6	3	6	6	5	7	5	4	2	38	74
5	2	2	6	2	4	9	5	1	31	59
4	—	4	4	3	9	5	1	—	26	46
3	—	1	2	2	4	4	3	1	17	36
2	1	3	3	4	7	3	6	—	27	29
1	1	1	1	3	6	7	5	—	24	18
0 (Low)	—	—	2	3	6	5	5	—	21	8
Total frequency	26	26	30	31	50	41	35	10	249	
Cumulative per cent	10	21	33	45	65	82	96	100		
Midpoint of content percentiles	5	16	27	39	55	74	89	98		
Median of intensity percentiles	85	64	52	49	38	39	32	78		

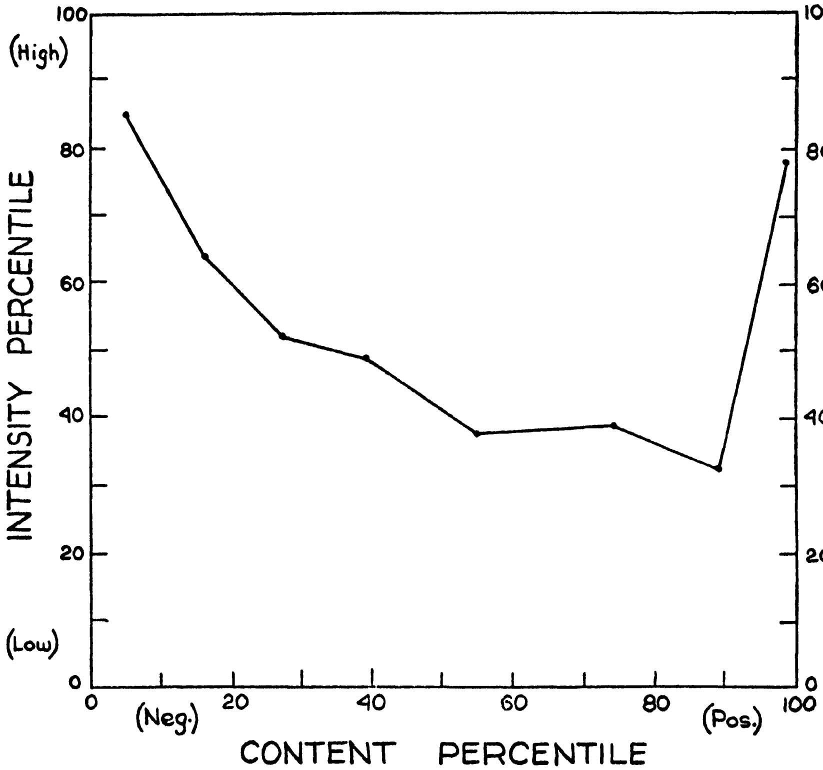


Figure 13. Satisfaction with one's job—educated privates.

The validity of the intensity function for analytical problems remains a field for future research. How much more the zero point can contribute to our knowledge of attitudes and opinions and how useful it can be for purposes of analysis has yet to be investigated. Briefly, we can say that flatbottom *U*-shaped curves indicate that there is not much difference between pros and cons; except for those relatively few people who are very extreme on either side, most of the people are relatively indifferent. Very sharp *V* curves indicate a clear distinction between being positive and being negative.

The intensity function does appear to offer a promising solution to the problem of obtaining a meaningful cutting point in a universe of content, such that those persons whose rank order is on one side of the cutting point can be said to be positive in their expression of attitude and people on the other side negative. It requires no a priori judgment, is independent of the sample of questions used, of the way they are worded and the like, and is completely objective.

Procedure for Intensity Measurement

A great deal of experimentation with different methods of asking intensity of feeling has indicated that intensity can be measured by asking after each attitude question about strength of feeling (i.e., "How strongly do you feel about this?") or degree of certainty (i.e., "How sure are you of your answer?") or difficulty of choice (i.e., "How hard was it for you to answer this question?"), or by mixing all three forms of the intensity question. Matched groups, answering the same content questions but different intensity questions, show no significant differences in the intensity curve or the zero range. All forms of the intensity question produce quasi scales of intensity.

Determining Both Content and Intensity by the Same Question

The method of using two parts to a question to determine content and intensity involves almost doubling the part of the questionnaire devoted to a particular area. It would be desirable, if possible, to devise a means of determining an intensity function without increasing the length of the questionnaire. It so happens that other research workers have often combined the intensity with the content without distinguishing between the two. In the Likert form of phrasing questions, for example, a declarative statement is followed

TABLE 14

ATTITUDE TOWARD THE ARMY

QUESTION: "When the Army score card plan gets going, do you think it will really be carried out in the way it is supposed to be?"

For those men who answer "Yes"

(Content Scores by Intensity Scores)

Intensity rank	*Content rank* (*Neg.*) *0*	*1*	*2*	*3*	*4*	*5*	*6*	(*Pos.*) *7*	*Total freq.*	*Cum. %*
7 (High)	33	44	42	60	39	43	37	82	380	100
6	*12*	*30*	33	31	29	37	37	*33*	242	76
5	9	15	*20*	40	20	21	36	20	181	60
4	5	12	23	*36*	*25*	*28*	*23*	6	158	48
3	2	11	15	28	25	26	17	12	136	38
2	1	10	17	26	26	24	20	8	132	30
1	2	5	13	31	26	28	17	5	127	21
0 (Low)	3	13	21	41	30	38	46	10	202	13
Total frequency	67	140	184	293	220	245	233	176	1,558	
Cumulative per cent	4	13	25	44	58	74	89	100		
Midpoint of content percentiles	2	9	19	34	51	66	81	94		
Median of intensity percentiles	75	62	50	44	40	41	46	73		

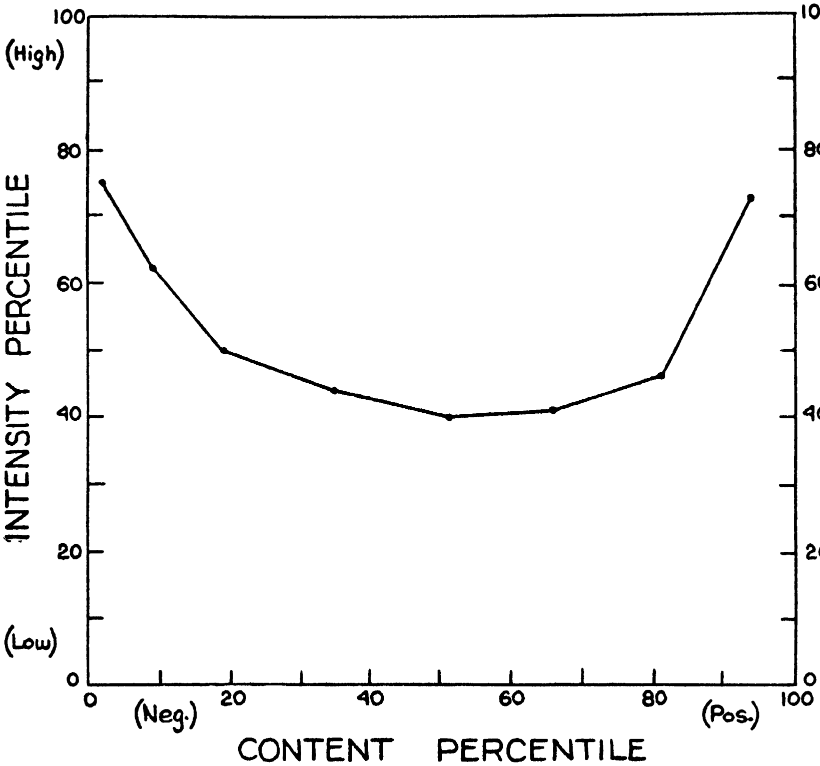


Figure 14. Attitude toward the Army (those who answered "yes" to the question).

TABLE 15

ATTITUDE TOWARD THE ARMY

QUESTION: "When the Army score card plan gets going, do you think it will really be carried out in the way it is supposed to be?"

For those men who answer "No"

(Content Scores by Intensity Scores)

Intensity rank	*Content rank* (*Neg.*) *0*	*1*	*2*	*3*	*4*	*5*	*6*	(*Pos.*) *7*	*Total freq.*	*Cum. %*
7 (High)	*126*	124	72	49	34	13	16	12	446	100
6	42	*61*	44	35	18	9	7	*9*	215	67
5	13	30	*36*	20	19	11	3	—	132	51
4	12	30	36	*33*	*17*	*9*	*9*	7	153	41
3	10	25	25	27	12	*11*	7	2	119	29
2	3	14	15	13	26	7	5	1	84	21
1	2	13	21	12	14	9	6	1	78	14
0 (Low)	8	19	15	26	23	15	6	1	113	8
Total frequency	216	306	264	215	163	84	59	33	1,340	
Cumulative per cent	16	39	59	75	87	93	98	100		
Midpoint of content percentiles	8	28	49	67	81	90	95	99		
Median of intensity percentiles	71	58	46	40	34	29	36	59		

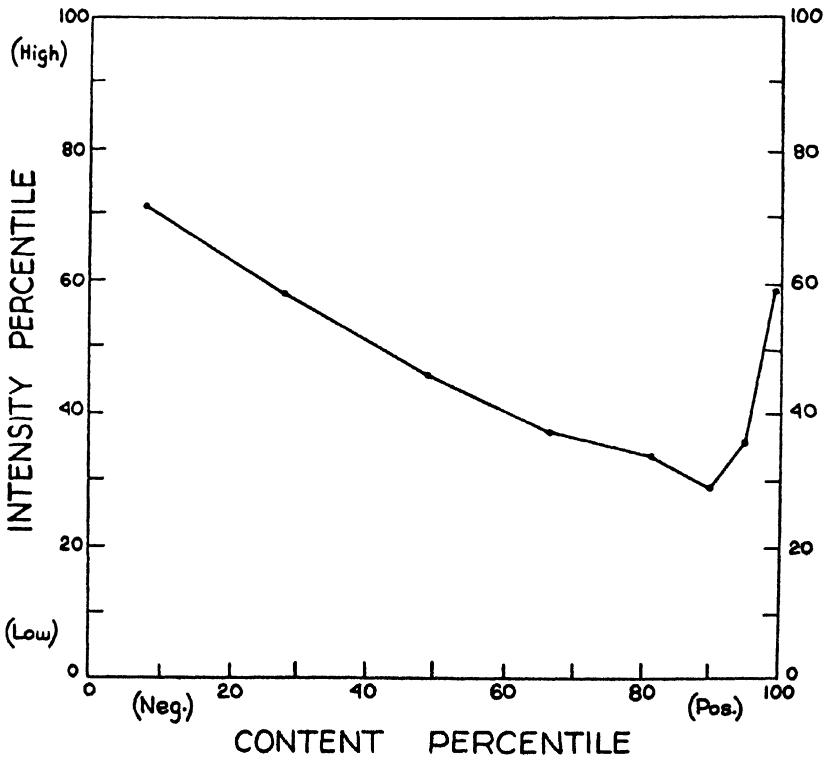


Figure 15. Attitude toward the Army (those who answered "no" to the question).

by an intensity check list. According to this, an attitude question would be put as follows:

Do you agree or disagree with the following statement:

Being a WAC is bad for a girl's reputation.

(Check one)

______ Strongly agree
______ Agree
______ Undecided
______ Disagree
______ Strongly disagree

Such a form contains both the content and intensity aspect simultaneously. If these can now be separated empirically, keeping the question in the same form, we could obtain the desired two scores—content and intensity—without lengthening the questionnaire.[8]

Using the universe of attitude toward the WAC, a series of questions containing five-part answer categories was asked. These questions were also followed by the intensity question, "How sure are you of your answer?" in order to permit a direct comparison of the two different methods of measuring intensity.[9] To obtain the content score, the questions were scaled in the ordinary fashion, using all five categories of the question. To obtain an intensity score, the answer categories were "folded over," that is, the most extreme "positive" category was combined with the most extreme "negative" category to denote strong intensity. The middle three categories were combined to signify mild intensity. In this way the respondent could be given an intensity score for each item.

In addition to this intensity score determined by "folding over" a single question, the respondent was given an intensity score based upon the separate intensity questions. Finally he was given a third intensity score based upon both his "fold-over" score and his intensity score from the separate intensity question. These three

[8] An immediate objection to such an attempt is, of course, that obtaining two scores from the same question implies that the scores are not experimentally independent. Using the same data twice may introduce some form of spuriousness, because experimental errors will be correlated in the two scores.

[9] The question might be asked, "What effect does the following up of each content question with an intensity question have upon the responses to the content question?" To investigate this problem, the area of attitude toward officers was studied. In one questionnaire the content questions were asked, as previously described, with a part *b* of intensity following each content question. In the second questionnaire only the content questions were asked. The reproducibility of the content scales in both cases was about .87. The scalogram pictures of both were very similar. The marginals on the questions were then, of course, also quite similar, indicating that the inclusion of the part *b* on intensity did not affect the men's answers to the part *a* on content.

intensity scores were then cross tabulated with his content scale score.

Figures 16, 17, and 18 (based on Tables 16, 17, and 18) show the results obtained by these three methods. The intensity curves and the zero points coincide almost perfectly. According to all three techniques the ratio of unfavorable to favorable opinion about the WAC is 3 to 1.

From this result it would appear that it is possible to measure *intensity of feeling* along with *direction of content* in a single question. However, additional experiments in the universes of attitude toward officers and satisfaction with one's job failed to produce the desired intensity curves by means of the "fold-over" technique, *although these curves were obtained by means of a separate intensity question.* While further research may serve to improve the measurement of content and intensity by means of a single question, for the present it would seem safer to use separate intensity questions. This latter method has the advantage of calling for two distinct decisions from the respondent in regard to content and intensity separately and thus permitting the correlation of two independently determined scores.

Generalized Intensity

The technique of asking "How strongly do you feel about this?" after each question is admittedly a crude one and accounts for much of the error in the observed relationship between intensity and content. If the intrinsic intensity could be ascertained, the relationship should be perfect. How error is introduced by the technique just described can be shown in several ways. Some men would say "undecided" to a question. When asked part *b* about strength of feeling, they would say they felt "very strongly." When they were asked why they said "very strongly" to part *b* when they were undecided on part *a*, they would answer to the effect that by "very strongly" they meant that the problem was very important, or else that they were very sure that they were undecided, or some such thing. Thus, many of the responses were out of context and contributed to error in the frequency distribution.

Even if all the responses were in context, there is still a contribution to error from the fact that the degrees of meaning of the words vary from subgroup to subgroup of the population. Verbal habits of people are considerably different. Some people will say "strongly agree" to almost anything when they are in favor of it,

TABLE 16

ATTITUDE TOWARD THE WAC
Content and Intensity Measured by Two Separate Questions
(Content Scores by Intensity Scores)

Intensity rank	*Content rank* (*Neg.*) *0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	(*Pos.*) *8*	*Total freq.*	*Cum. %*
6 (High)	*29*	63	78	47	21	21	13	12	*8*	292	100
5	8	*29*	50	34	21	21	6	7	*2*	178	83
4	—	17	*60*	46	22	28	11	7	3	184	73
3	3	23	27	*65*	39	36	13	*15*	—	221	62
2	2	2	22	65	*60*	60	*27*	17	1	256	49
1	—	4	20	48	45	*123*	34	9	1	284	35
0 (Low)	—	3	12	61	59	146	30	3	1	315	18
Total frequency	42	141	259	366	267	435	134	70	16	1,730	
Cumulative per cent	2	11	26	47	62	87	95	99	100		
Midpoint of content percentiles	1	7	18	36	54	75	91	97	100		
Median of intensity percentiles	88	80	73	51	42	28	36	55	83		

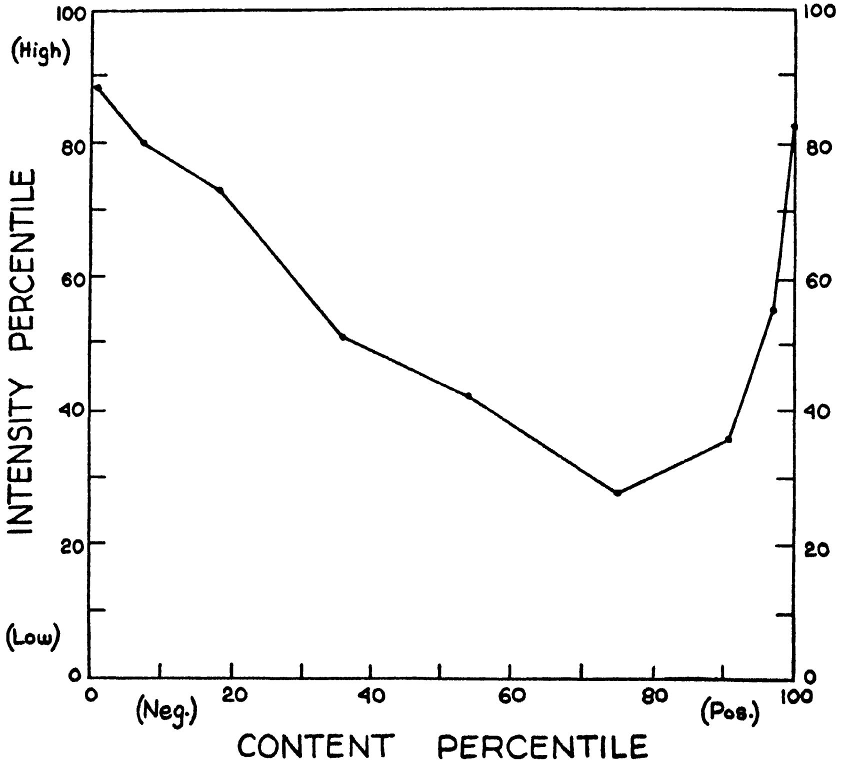


Figure 16. Attitude toward the WAC (content and intensity measured by two separate questions).

TABLE 17

ATTITUDE TOWARD THE WAC
Content and Intensity Measured by the Same Question
(Content Scores by Intensity Scores)

Intensity rank	*Content rank* (*Neg.*) *0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	(*Pos.*) *8*	*Total freq.*	*Cum. %*
6 (High)	*27*	13	—	—	1	2	1	—	2	46	100
5	11	*67*	19	1	2	3	1	5	4	113	97
4	1	48	94	26	10	7	9	15	*6*	216	91
3	1	10	*110*	104	33	12	29	*22*	2	323	78
2	—	1	29	*157*	76	64	*32*	20	2	381	60
1	2	1	4	72	*94*	*177*	28	8	—	386	38
0 (Low)	—	1	3	6	51	170	34	—	—	265	15
Total frequency	42	141	259	366	267	435	134	70	16	1,730	
Cumulative per cent	2	11	26	47	62	87	95	99	100		
Midpoint of content percentiles	1	7	18	36	54	75	91	97	100		
Median of intensity percentiles	98	92	76	52	35	21	41	66	87		

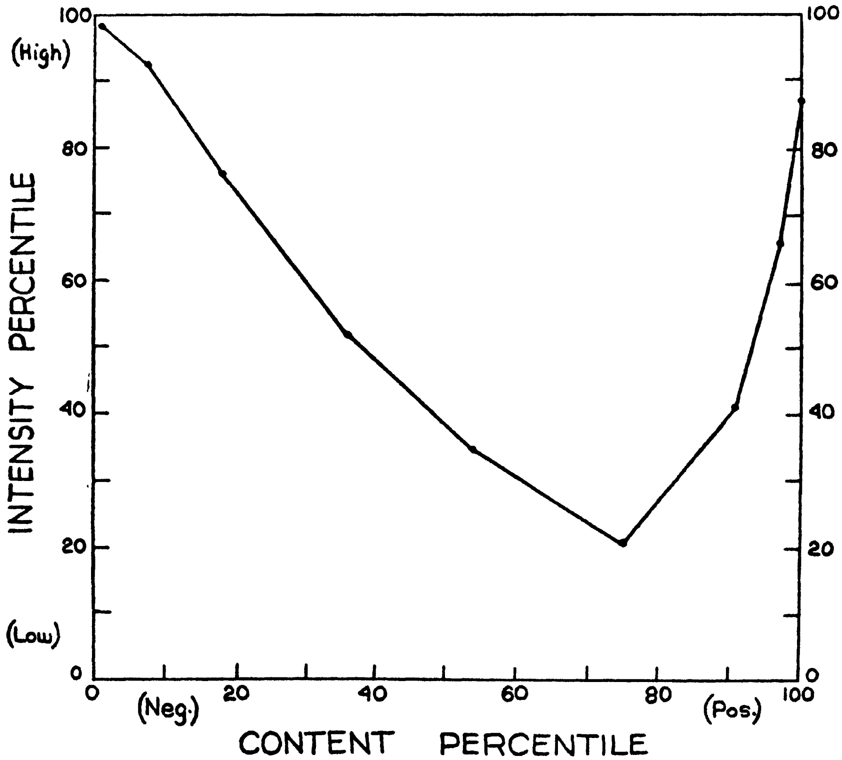


Figure 17. Attitude toward the WAC (content and intensity measured by the same question).

where other people would say "agree" under the same circumstances. Especially with respect to the intensity questions, there are people who say "very strongly" to every question. This tendency to use or not to use strong adjectives we shall call "generalized verbal intensity." That such a concept exists as a quasi scale for the Army population has been shown in several surveys.

TABLE 18

Attitude toward the WAC
Intensity Measured by Both Single and Separate Questions
(Content Scores by Intensity Scores)

Intensity rank	*Content rank* (*Neg.*) *0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	(*Pos.*) *8*	*Total freq.*	*Cum. %*
15 (High)	*28*	13	—	—	—	—	1	1	4	47	100
14	3	38	16	1	1	2	—	1	2	64	97
13	1	*31*	44	11	2	3	5	4	*2*	103	94
12	5	20	34	28	10	5	7	10	*2*	121	88
11	—	13	*48*	38	14	10	6	6	2	137	81
10	3	10	30	28	16	21	9	9	2	128	73
9	1	5	23	49	20	13	7	*15*	—	133	65
8	—	4	29	*48*	40	32	19	8	—	180	58
7	—	2	12	38	*33*	41	*15*	9	1	151	47
6	—	2	10	48	31	57	18	5	1	172	38
5	—	1	4	40	39	*70*	15	2	—	171	29
4	1	—	3	18	21	85	14	—	—	142	19
3	—	—	1	13	15	50	9	—	—	88	10
2	—	1	2	5	15	29	6	—	—	58	5
1	—	—	2	1	8	11	2	—	—	24	2
0 (Low)	—	1	1	—	2	6	1	—	—	11	1
Total frequency	42	141	259	366	267	435	134	70	16	1,730	
Cumulative per cent	2	11	26	47	62	87	95	99	100		
Midpoint of content percentiles	1	7	18	36	54	75	91	97	100		
Median of intensity percentiles	98	90	75	52	39	24	40	63	88		

Generalized verbal intensity can be measured by studying several different universes of content (attitudes) in a single survey. One part *b* question on intensity is selected from each of the different areas at random. Scalogram 14 shows the different content questions included and the pattern of intensity responses to these questions. From this scalogram it can be seen that the eleven random intensity questions form a quasi scale. It is therefore possible to

rank individuals along a scale dimension of what we have called "generalized verbal intensity."

The manner in which this generalized verbal intensity contributes to the errors in the present method of measuring intensity can be clearly demonstrated by the following procedure. Respondents are divided into three approximately equal groups depending upon their

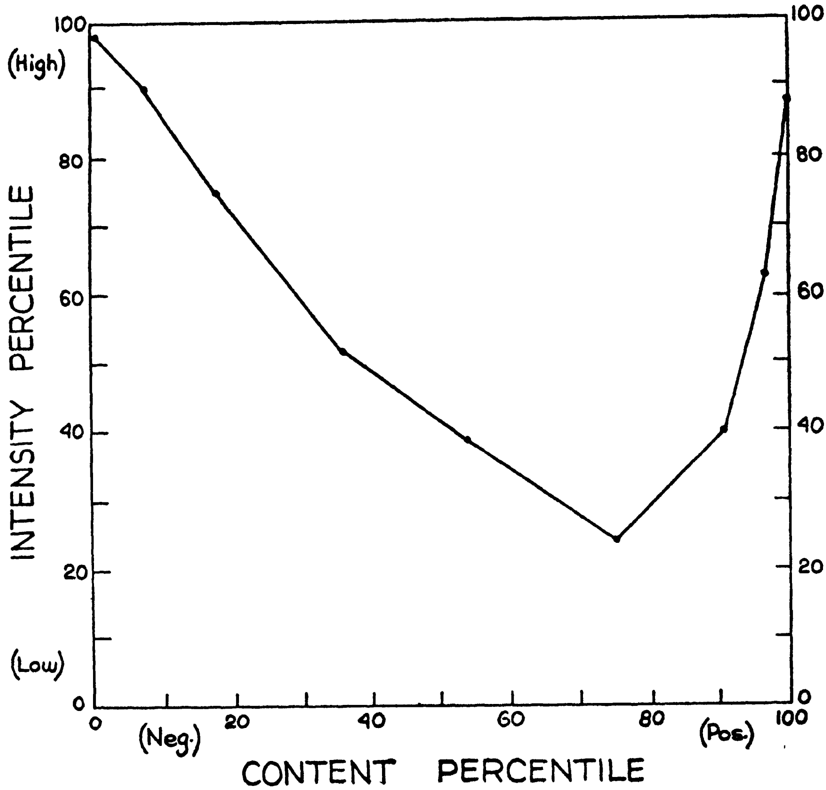
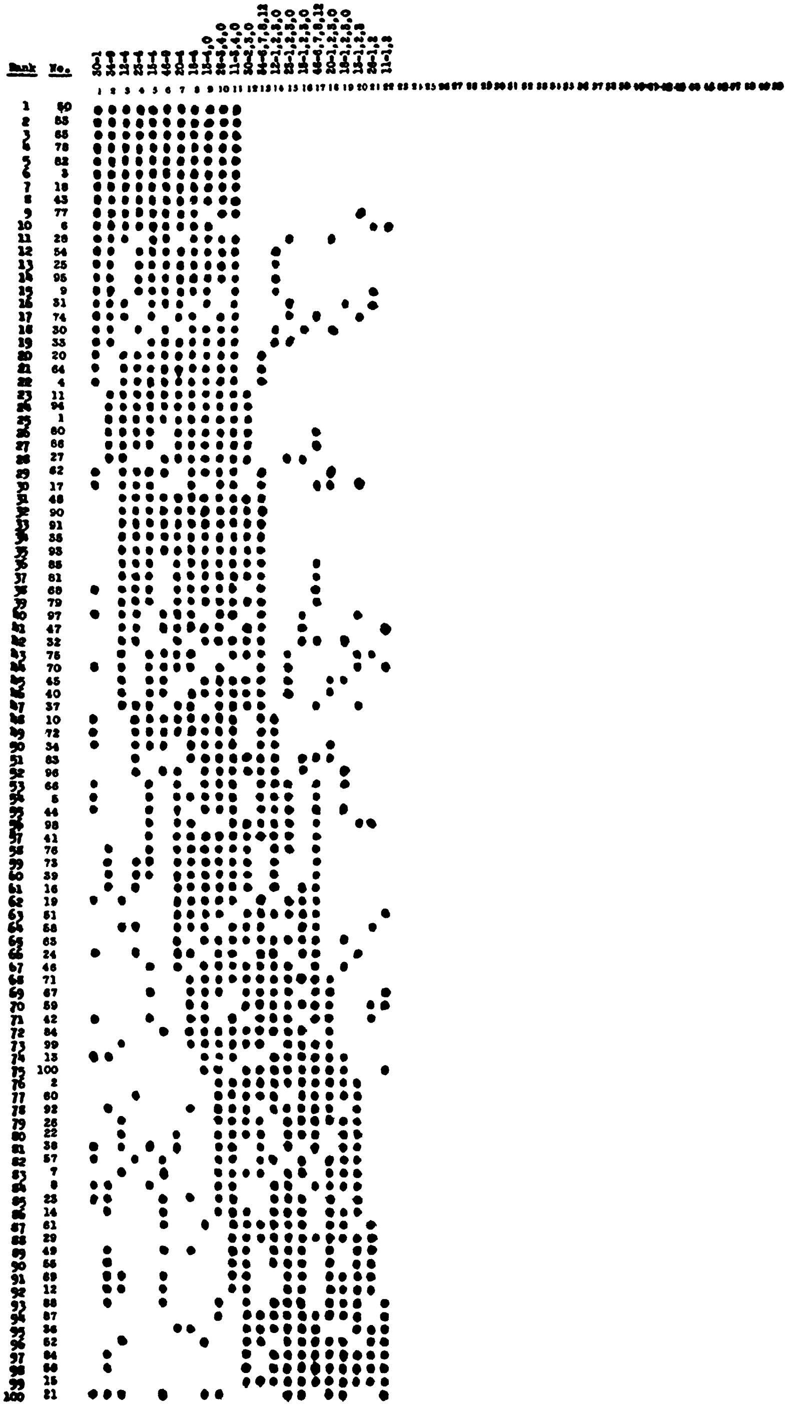


Figure 18. Attitude toward the WAC (intensity measured by both single and separate questions).

position on the scale of generalized verbal intensity. Those persons in the upper third are individuals who tend to give strong intensity answers, those in the middle third tend to give moderate intensity answers, while those in the lower third tend to give weak intensity answers. Using the area of attitude toward the WAC, we then compute separate cross tabulations of content scores by intensity scores for each of these three generalized intensity groups.

SCALOGRAM No. 14. GENERALIZED VERBAL INTENSITY

Respondent order *Category order*



Scalogram No. 14. Generalized Verbal Intensity

Questions and Answer Categories

11a. Considering their responsibilities, how do you feel about the privileges that officers get compared with those which enlisted men get?

b. How strongly do you feel about this?

1 ______ Not strongly at all
2 ______ Not so strongly
3 ______ Fairly strongly
4 ______ Very strongly
0 ______ No answer

12a. On the whole, how is your morale?

b. Same as 11b.

13a. All things considered, do you think the Army is run about as well as possible, or do you think it could be run better?

b. Same as 11b.

15a. In general, do you feel you yourself have gotten a square deal from the Army?

b. Same as 11b.

18a. On the whole, do you think the Army gives a man a chance to show what he can do?

b. Same as 11b.

20a. What do you think of discipline in the Army?

b. Same as 11b.

23a. Did your officers use their rank in ways that seemed unnecessary to you?

b. Same as 11b.

26a. The Army has set up an Inspector-General system (the IG or Air Inspector) to handle complaints from enlisted men. How well do you think this system actually works?

b. Same as 11b.

30a. In your opinion, is the Army Score Card (the point system) being carried out the way it was supposed to be?

b. How strongly do you feel about this?

1 ______ Very strongly
2 ______ Fairly strongly
3 ______ Not strongly
0 ______ No answer

34a. Here are some statements about Army rules and regulations and their enforcement. For each statement, check the answer that best shows how closely you agree with it.

A person should obey only those Army rules and regulations which seem reasonable.

b. How hard was it for you to answer this question?

6 ______ Very hard
7 ______ Fairly hard
8 ______ Not so hard
9 ______ Not hard at all
12 ______ No answer

46a. With which one of these statements concerning postwar relations with Russia do you come closest to agreeing?

b. How strongly do you feel about this?

6 ______ Not at all strongly
7 ______ Not so strongly
8 ______ Fairly strongly
9 ______ Very strongly
12 ______ No answer

The results of these cross tabulations are shown in Tables 19, 20, and 21. By looking at the location of the medians of intensity for each content rank, we can see quite clearly why there is so much spread in the combined cross tabulation of all three intensity groups. The curve of medians for persons with strong generalized intensity (Table 19) is concentrated in the upper intensity ranks of the specific area being studied—attitude toward the WAC. The curve of medians for persons with weak generalized intensity is located in the lower intensity ranks, while the curve of medians for persons with moderate generalized intensity falls within the middle intensity ranks. Combining all three generalized intensity groups for this attitude area serves to spread out the intensity scores through all intensity ranks, although the median intensity score for any single

TABLE 19

ATTITUDE TOWARD THE WAC
High Generalized Intensity
(Content Scores by Intensity Scores)

Intensity rank	*Content rank* (*Neg.*) *0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	(*Pos.*) *8*	*Total freq.*	*Cum. %*
15 (High)	*20*	11	—	—	—	—	—	1	4	36	100
14	1	*31*	14	—	1	2	—	1	*2*	52	93
13	—	16	37	9	1	1	5	3	2	74	83
12	2	8	*25*	21	10	1	5	8	1	81	69
11	—	1	18	19	11	3	3	*3*	1	*59*	53
10	1	—	9	*12*	8	15	*2*	2	—	49	42
9	1	1	6	13	*5*	6	*2*	5	—	39	33
8	—	1	1	11	*10*	*13*	3	3	—	42	25
7	—	—	1	8	8	6	4	1	—	28	17
6	—	—	2	7	6	6	3	—	—	24	12
5	—	—	—	2	7	7	1	—	—	17	7
4	1	—	—	2	2	6	1	—	—	12	4
3	—	—	—	—	2	2	—	—	—	4	1
2	—	—	—	—	1	1	1	—	—	3	1
1	—	—	—	—	—	—	—	—	—	—	—
0 (Low)	—	—	—	—	—	—	—	—	—	—	—
Total frequency	26	69	113	104	72	69	30	27	10	520	
Cumulative per cent	5	18	40	60	74	87	93	98	100		
Midpoint of content percentiles	3	12	29	50	67	80	90	95	99		
Median of intensity percentiles	96	86	65	40	25	21	33	51	88		

content rank retains its relative position to median intensity scores for the other content ranks. This spread of cases in intensity ranks other than the rank of median intensity is what we call errors of intensity measurement.

It is significant to note that computing content and intensity percentiles and drawing the intensity curve for each generalized intensity group separately (Figures 19, 20, and 21) takes into account the

TABLE 20

ATTITUDE TOWARD THE WAC
Medium Generalized Intensity
(Content Scores by Intensity Scores)

Intensity rank	*Content rank* (*Neg.*) *0*	*1*	*2*	*3*	*4*	*5*	*6*	(*Pos.*) *7–8*	*Total freq.*	*Cum. per cent*
15 (High)	*8*	2	—	—	—	—	1	—	11	100
14	2	6	2	1	—	—	—	—	11	98
13	—	11	9	2	1	2	—	1	26	96
12	2	*7*	16	6	—	4	2	1	38	92
11	—	7	16	15	3	6	2	4	53	86
10	1	8	*16*	11	6	4	5	4	55	77
9	—	4	19	27	8	5	3	*6*	72	68
8	—	2	12	*25*	22	11	8	1	*81*	56
7	—	—	9	17	*16*	21	*4*	6	73	43
6	—	—	4	19	12	*22*	8	—	65	31
5	—	1	2	13	10	15	6	1	48	20
4	—	—	—	7	11	19	2	—	39	12
3	—	—	—	4	6	9	4	—	23	6
2	—	—	—	—	2	5	1	—	8	2
1	—	—	—	—	2	1	1	—	4	1
0 (Low)	—	—	—	—	—	—	1	—	1	—
Total frequency	13	48	105	147	99	124	48	24	608	
Cumulative per cent	2	10	27	51	68	88	96	100		
Midpoint of content percentiles	1	6	19	39	60	78	92	98		
Median of intensity percentiles	99	88	72	50	36	27	34	64		

"bias" of generalized intensity and produces almost exactly the same *U* curve and zero point for all three groups. The similarity of the intensity function for the different generalized intensity groups indicates the important property of the invariance of the intensity function. The above results were also obtained in other attitude universes studied separately for the different generalized verbal intensity groups.

Generalized Verbal Intensity and the Single Question

It was also found that responses to individual questions were related to this generalized intensity. This finding emphasizes all the more that responses to any single content question must be regarded with caution and reemphasizes the need for an objective technique for obtaining a zero point such as that described in this chapter.

TABLE 21

ATTITUDE TOWARD THE WAC
Low Generalized Intensity
(Content Scores by Intensity Scores)

Intensity rank	*Content rank* (*Neg.*) *0–1*	*2*	*3*	*4*	*5*	*6*	(*Pos.*) *7–8*	*Total freq.*	*Cum. %*
15 (High)	1	—	—	—	—	—	—	1	100
14	1	—	—	—	—	—	—	1	100
13	5	1	—	—	—	—	—	6	100
12	6	2	1	—	—	—	2	11	99
11	5	6	4	—	1	1	—	17	97
10	3	5	5	2	2	2	5	24	94
9	—	6	11	7	2	2	4	32	90
8	1	7	10	8	8	8	4	46	85
7	2	1	13	9	14	7	3	49	77
6	2	3	22	13	29	7	6	82	69
5	—	1	25	22	48	8	1	105	55
4	—	3	9	8	60	11	—	91	38
3	—	1	9	7	39	5	—	61	23
2	1	2	5	12	23	4	—	47	13
1	—	2	1	6	10	1	—	20	5
0 (Low)	1	1	—	2	6	—	—	10	2
Total frequency	28	41	115	96	242	56	25	603	
Cumulative per cent	5	11	31	46	87	96	100		
Midpoint of content percentiles	2	8	21	38	67	91	98		
Median of intensity percentiles	96	84	61	48	34	53	82		

An example of how "generalized" intensity can be related to responses to a particular content question is the following—the sample of soldiers was divided into three groups according to their generalized intensity: low, medium, and high. When asked the question:

> All things considered, do you think the Army is run about as efficiently as possible, or do you think it could be run better?

the responses for the three groups turned out to be:

	GENERAL INTENSITY		
	Low	*Medium*	*High*
It is run as well as possible	38%	32%	28%
It could be run somewhat better	42	36	27
It could be run a lot better	20	32	45
Total per cent	100%	100%	100%
Total cases	*441*	*800*	*525*

The categories obviously mean different things to the different kinds of people.

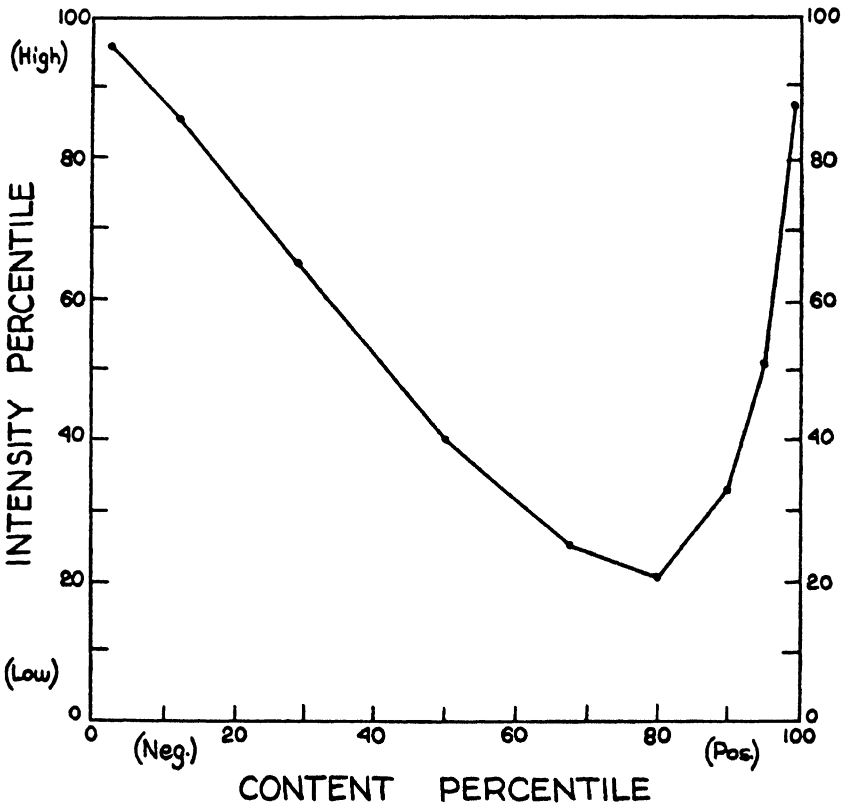


Figure 19. Attitude toward the WAC (high generalized intensity).

Another example would be the question:

On the whole, would you say you are well adjusted or poorly adjusted to Army life?

	GENERAL INTENSITY		
	Low	*Medium*	*High*
Very well adjusted	8%	17%	29%
Fairly well adjusted	56	47	34
Not so well adjusted	23	18	14
Very poorly adjusted	13	18	23
Total per cent	100%	100%	100%
Total cases	*441*	*800*	*525*

The percentage of respondents answering "very well adjusted" varied from 8 per cent to 29 per cent depending upon the generalized verbal intensity habits of the respondent.

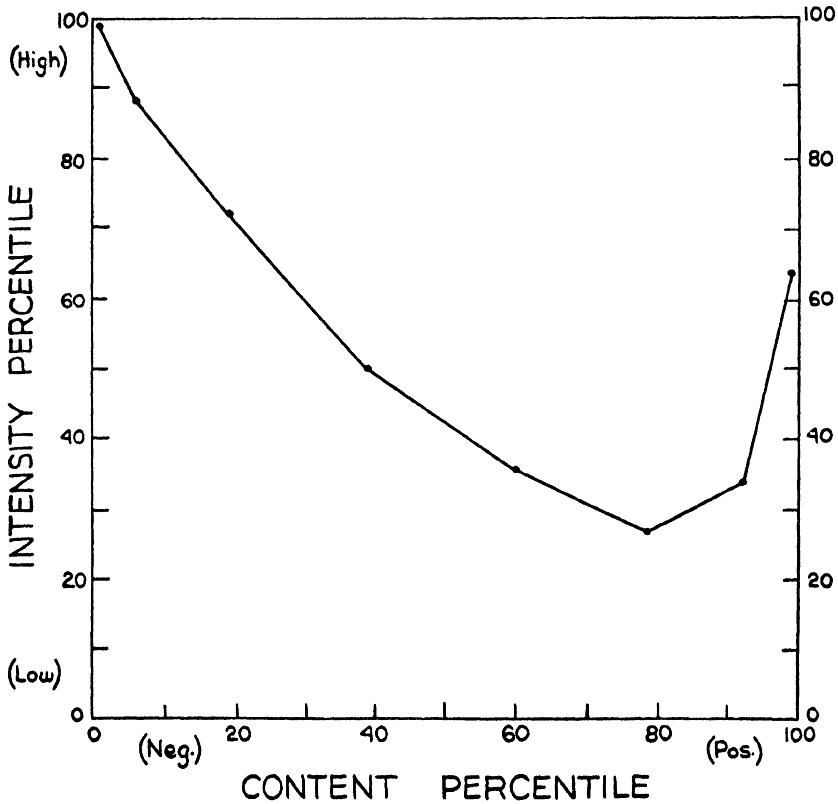


Figure 20. Attitude toward the WAC (medium generalized intensity).

The distribution of content scale scores was affected as well as the single question marginals. For example, the scale of attitude toward the Army showed the following relationship to generalized intensity:

Scale score	GENERAL INTENSITY *Low*	*Medium*	*High*
0 to 6	18%	35%	45%
7 to 9	44	32	23
10 to 16	38	33	32
Total per cent	100%	100%	100%
Total cases	*441*	*800*	*525*

A similar relationship between generalized verbal intensity and single question marginal frequencies or scale frequencies was ob-

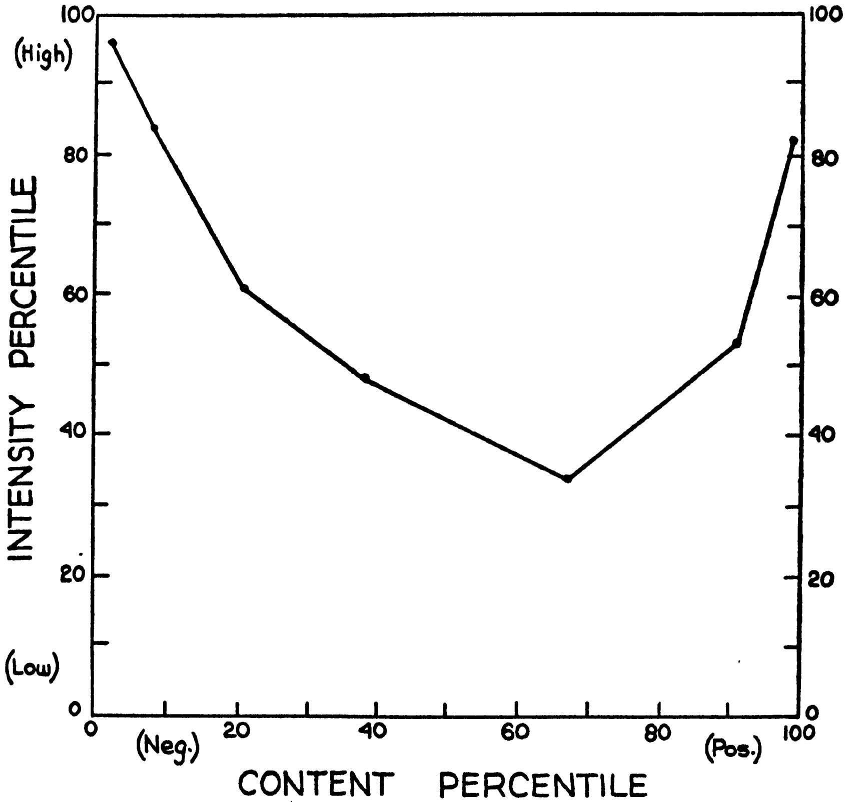


Figure 21. Attitude toward the WAC (low generalized intensity).

served in almost all cases. Responses to questions, therefore, involve not only the attitudes and opinions of the individual but also his verbal habits of expression.

Empirical Demonstration of the Invariance of the Zero Point

How can the claim be tested that the proposed method produces results independent of any "biases" in the specific questions asked? Perhaps the best way to destroy this claim would be to take two series of opinion questions on the same topic, but "biased" in completely opposite directions and producing apparently widely different results, and then to see if the inclusion of the intensity measurement serves to "correct" these "biases" and produce the same final result for both sets of questions. Stated in other words, let us suppose that two different polling agencies wanted to find out how many people hold "favorable" or "unfavorable" opinions concerning a certain issue. One agency asks several questions which indicate that *less* than half of the people are "favorable" on each of the questions asked. The other agency asks several different questions which indicate, however, that *more* than half of the people are "favorable" on each of the questions that it asked. Both sets of questions deal with the same issue, but apparently produce completely opposite results. Here is the problem of question "bias" stated in its most extreme terms. What kind of solution does the present method of intensity analysis offer?

TABLE 22

COMPARISON OF DISTRIBUTION OF REPLIES TO "UNFAVORABLY BIASED" AND "FAVORABLY BIASED" SERIES OF QUESTIONS

SERIES	OPINION OF ARMY *Question number*	*Per cent favorable*	OPINION OF OFFICERS *Question number*	*Per cent favorable*
Series "A" ("Unfavorably biased")	1	8%	1	7%
	2	14	2	15
	3	16	3	23
	4	31	4	29
	5	33	5	32
	6	43	6	47
Series "B" ("Favorably biased")	7	58	7	50
	8	65	8	52
	9	70	9	68
	10	79	10	75
	11	81	11	83
	12	90	12	93

List 1

Attitude toward the Army

(The manner in which the answer categories were dichotomized can be seen from the weights of 0 or 1 appearing before each category.)

Series A

("Unfavorably Biased" Dichotomization)

1. On the whole, do you think the Army gives a man a chance to show what he can do?
   (1) A very good chance
   (0) A fairly good chance
   (0) Not much of a chance
   (0) No chance at all
   (0) Undecided

2. In general, how well do you think the Army is run?
   (1) It is run very well
   (0) It is run pretty well
   (0) It is not run so well
   (0) It is run very poorly
   (0) Undecided

3. Do you think when you are discharged you will go back to civilian life with a favorable or unfavorable attitude toward the Army?
   (1) Very favorable
   (0) Fairly favorable
   (0) About 50–50
   (0) Fairly unfavorable
   (0) Very unfavorable

4. In general, do you think the Army has tried its best to see that men get as square a deal as possible?
   (1) Yes, it has tried its best
   (0) It has tried some but not hard enough
   (0) It has hardly tried at all

5. Do too many of the things you have to do in the Army seem unnecessary?
   (1) No, not too many of them seem unnecessary
   (0) Yes, too many of them seem unnecessary

6. In the Army, some jobs are naturally harder and more dangerous than others and the Army has to put men where it thinks they are needed.
   Considering everything, do you think the Army is trying its best to see that, as far as possible, no man gets more than his fair share of the hard and dangerous jobs?
   (1) Yes, it is trying its best
   (0) It is trying some, but not hard enough
   (0) It is hardly trying at all

Series B

("Favorably Biased" Dichotomization)

7. In general, how interested do you think the Army is in your welfare?
   (1) Very much
   (1) Pretty much
   (0) Not so much
   (0) Not at all

8. What do you think of the statement that *"The Army makes a man out of you?"*
   (1) There's a lot to it
   (1) There may be something to it, but I'm still doubtful
   (0) There is not much to it

9. All things considered, do you think the Army is run about as efficiently as possible, or do you think it could be run better?
   (1) It is run about as well as possible, everything considered
   (1) It could be run somewhat better
   (0) It could be run a lot better

10. Do you think the Army is trying its best to see that the men who have the hard and dangerous jobs get the special consideration and breaks they deserve?
    (1) Yes, it is trying its best
    (1) It is trying some, but not hard enough
    (0) It is hardly trying at all

11. In general, do you feel you yourself have gotten a square deal from the Army?
    (1) Yes, in most ways I have
    (1) In some ways, yes, in other ways, no
    (0) No, on the whole I haven't gotten a square deal

12. Do you feel that the Army is trying its best to look out for the welfare of enlisted men?
    (1) Yes, it is trying its best
    (1) It is trying some, but not hard enough
    (0) It is hardly trying at all

List 2

Attitude toward Officers

(The manner in which the answer categories were dichotomized can be seen from the weights of 0 or 1 appearing before each category.)

Series A

("Unfavorably Biased" Dichotomization)

1. How much did you personally like your officers?
   (1) Very much
   (0) Pretty much
   (0) Not so much
   (0) Not at all

2. How do you feel about the privileges that officers get compared with those which enlisted men get?
   (0) Officers have *far too many* privileges
   (0) Officers have *a few too many* privileges
   (1) Officers have *about the right number* of privileges
   (1) Officers have *too few* privileges

3. How much did you personally respect your officers?
   (1) Very much
   (0) Pretty much
   (0) Not so much
   (0) Not at all

4. When you did a particularly good job did you usually get recognition or praise for it from your officers?
   (1) Always
   (1) Usually
   (0) Rarely
   (0) Never

5. How did you feel about the officers that had been selected by the Army?
   (1) They were the best ones that could have been picked
   (1) They were as good as any that could have been picked
   (1) Undecided
   (0) Somewhat better ones could have been picked
   (0) Much better ones could have been picked

6. Do you think that your officers generally did what they could to help you?
   (1) Yes, all the time
   (1) Yes, most of the time
   (0) No, they often did not
   (0) No, they almost never did

Series B

("Favorably Biased" Dichotomization)

7. How well do you feel that your officers understood your problems and needs?
   (1) They were very much aware of my problems and needs
   (1) They were fairly well aware of my problems and needs
   (0) They did not know very much about my real problems and needs

8. Do you feel that your officers recognized your abilities and what you were able to do?
   (1) Yes, I'm sure they did
   (1) Yes, I think they did, but I'm not sure
   (1) Undecided
   (0) No, I don't think they did

9. How many of your officers used their rank in ways that seemed unnecessary to you?
   (0) Almost all of them
   (0) Most of them
   (1) Some of them
   (1) Only a few of them
   (1) None of them

10. In general, how good would you say your officers were?
   (1) Very good
   (1) Fairly good
   (1) About average
   (0) Pretty poor
   (0) Very poor

11. Did your officers give you a chance to ask questions as to the reason why things were done the way they were?
   (1) Yes, always
   (1) Yes, usually
   (1) Undecided
   (0) No, not very often
   (0) No, almost never

12. How many of your officers took a personal interest in their men?
   (1) All of them
   (1) Most of them
   (1) About half of them
   (0) Few of them
   (0) None of them

The answer is clearly demonstrated by two examples of exactly the above nature. The first concerns the universe of attitudes of soldiers toward the Army, while the second concerns soldiers' opin-

ions of their officers. In each example a cross section of enlisted men in the United States was asked two sets of six questions each.[10] These two sets of questions produced widely opposing percentages of the population with "more favorable" and "less favorable" opinions on each question separately. The extent of this divergence can be seen from Table 22. Series A was composed of questions which resulted in *less* than half of the respondents expressing "favorable" attitudes *on all six questions,* while Series B was composed of questions which resulted in *more* than half of the respondents expressing "favorable" attitudes *on all six questions.* Each series of six questions was tested for scalability to assure that, from the respondent's point of view, they all concerned the same opinion issue.[11] The exact questions asked are given in Lists 1 and 2. The manner in which the answer categories were dichotomized can be seen from the scale weights of 0 or 1 given to each category.[12] Each opinion question was followed by an intensity question asking, "How strongly do you feel about this?"

From Table 22 it can be seen that, depending upon which series of six questions was used, soldiers' opinions of the Army and of their officers could be characterized as "favorable" or "unfavorable." One polling agency using Series A would conclude that the men "disliked" the Army and their officers, while another agency using Series B would conclude that the men "liked" the Army and their officers. However, our claim is that, were both agencies to ask intensity of feeling also and to correlate opinion scores by intensity scores, they would draw substantially the same conclusions, regardless of which set of questions was used. The *U*- or *J*-shaped curve obtained by plotting opinion scores against intensity scores would divide the population into the same percentages with "favorable" and "unfavorable" opinions, regardless of question "bias." What do our data show in this respect?

Tables 23a and 23b and 24a and 24b show the correlation between intensity scores and content scores for Series A and Series B. In order to plot these scores, the median intensity value is computed for each content value. These medians are italicized in the tables and

[10] The two examples, opinion of the Army and opinion of officers, were obtained from two separate surveys. In each case, the two sets of six questions were mixed at random so that the respondent answered twelve questions in each area without knowing which six went together.

[11] Scalograms for each of these two areas are given in preceding chapters.

[12] Since the answer categories represent only qualitative gradations from more to less "favorable," these weights can be given fairly arbitrarily to produce relatively high or low percentages with "favorable" opinions.

TABLE 23a

Attitude toward the Army
Series A ("Unfavorably Biased")

Intensity rank	*Content rank* (*Neg.*) *0*	*1*	*2*	*3*	*4*	*5*	(*Pos.*) *6*	*Total freq.*	*Cum. %*
6 (High)	222	112	49	45	26	18	*30*	502	100
5	*112*	61	29	30	18	*21*	5	276	73
4	*91*	*76*	38	31	*9*	5	5	255	58
3	76	61	*50*	*28*	11	8	2	236	44
2	58	48	47	17	16	1	1	188	31
1	50	50	33	29	5	4	—	171	21
0 (Low)	59	61	45	44	7	3	2	221	12
Total frequency	668	469	291	224	92	60	45	1,849	
Cumulative per cent	36	61	77	89	94	98	100		
Midpoint of content percentiles	18	49	69	83	92	96	99		
Median of intensity percentiles	58	47	37	41	55	64	80		

TABLE 23b

Attitude toward the Army
Series B ("Favorably Biased")

Intensity rank	*Content rank* (*Neg.*) *0*	*1*	*2*	*3*	*4*	*5*	(*Pos.*) *6*	*Total freq.*	*Cum. %*
6 (High)	*35*	*45*	63	70	96	90	164	563	100
5	1	18	*25*	*61*	42	68	105	320	70
4	—	11	21	29	*44*	*53*	*64*	222	52
3	—	3	9	31	35	56	72	206	40
2	1	—	11	18	34	58	68	190	29
1	—	2	5	12	24	49	75	167	19
0 (Low)	—	8	3	7	24	41	98	181	10
Total frequency	37	87	137	228	299	415	646	1,849	
Cumulative per cent	2	7	14	26	43	65	100		
Midpoint of content percentiles	1	4	10	20	35	54	83		
Median of intensity percentiles	84	71	66	57	49	41	42		

TABLE 24a

Attitude toward Officers
Series A ("Unfavorably Biased")

Intensity rank	*Content rank* (*Neg.*) *0*	*1*	*2*	*3*	*4*	*5*	(*Pos.*) *6*	*Total freq.*	*Cum. %*
6 (High)	266	105	56	34	14	15	12	502	100
5	161	102	53	34	20	15	*9*	394	83
4	*138*	113	66	43	20	*20*	10	410	69
3	123	*116*	78	37	24	13	4	395	55
2	91	96	*72*	*58*	*35*	6	1	359	41
1	81	112	103	68	27	12	2	405	29
0 (Low)	60	108	109	96	48	15	1	437	15
Total frequency	920	752	537	370	188	96	39	2,902	
Cumulative per cent	32	58	76	89	95	99	100		
Midpoint of content percentiles	16	45	67	83	92	97	99		
Median of intensity percentiles	66	48	39	33	36	56	71		

TABLE 24b

Attitude toward Officers
Series B ('Favorably Biased")

Intensity rank	*Content rank* (*Neg.*) *0*	*1*	*2*	*3*	*4*	*5*	(*Pos.*) *6*	*Total freq.*	*Cum. %*
6 (High)	*26*	*67*	*84*	91	94	80	68	510	100
5	5	22	*50*	80	75	69	54	355	82
4	4	14	43	*70*	99	89	64	383	70
3	—	6	24	67	*115*	110	74	396	57
2	—	8	16	53	103	*145*	*97*	422	43
1	—	5	12	31	87	129	80	344	29
0 (Low)	8	4	7	28	86	169	190	492	17
Total frequency	43	126	236	420	659	791	627	2,902	
Cumulative per cent	1	6	14	28	51	78	100		
Midpoint of content percentiles	1	4	10	21	40	65	89		
Median of intensity percentiles	85	83	74	63	50	39	35		

are plotted in Figures 23 and 24. These charts show quite clearly the effect of the "biased" selection of questions and demonstrate how this "bias" can be taken care of by means of the present method.

The agency using the "unfavorably biased" Series A on attitude

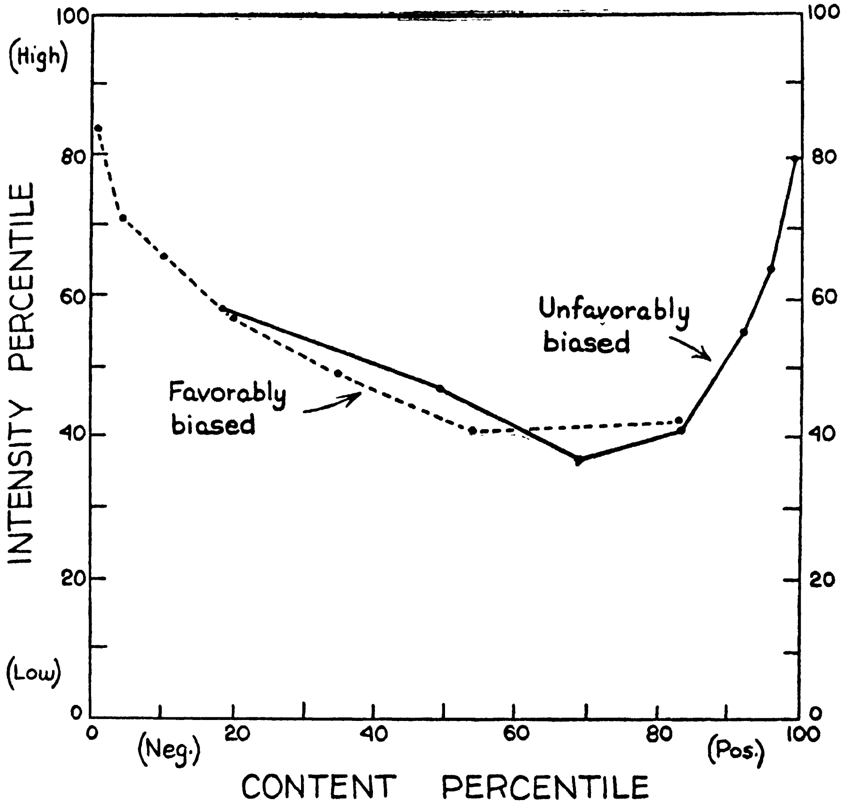


Figure 23*. Attitude toward the Army.
Series A—positive end of universe (Table 23a—"unfavorably biased").
Series B—negative end of universe (Table 23b—"favorably biased").

* There is no Figure 22. To avoid confusion, numbering of scalogram tables and figures is kept parallel throughout this chapter.

toward the Army would find from its intensity curve that the low point of the curve fell at 69 per cent, indicating approximately only 31 per cent "favorable" and 69 per cent "unfavorable" in their attitude toward the Army. The agency using the "favorably biased" Series B would have to conclude that, despite the fact that none of its six questions showed *less* than 58 per cent "favorable," the low point of the curve fell between 54 per cent and 83 per cent

(a crude single point estimate can be taken as the midpoint of this interval, or 68 per cent, indicating only 32 per cent "favorable"); this agency would see that, to determine this point more closely, questions showing greater "unfavorableness" toward the Army would have to be included in the series asked. Thus, both agencies

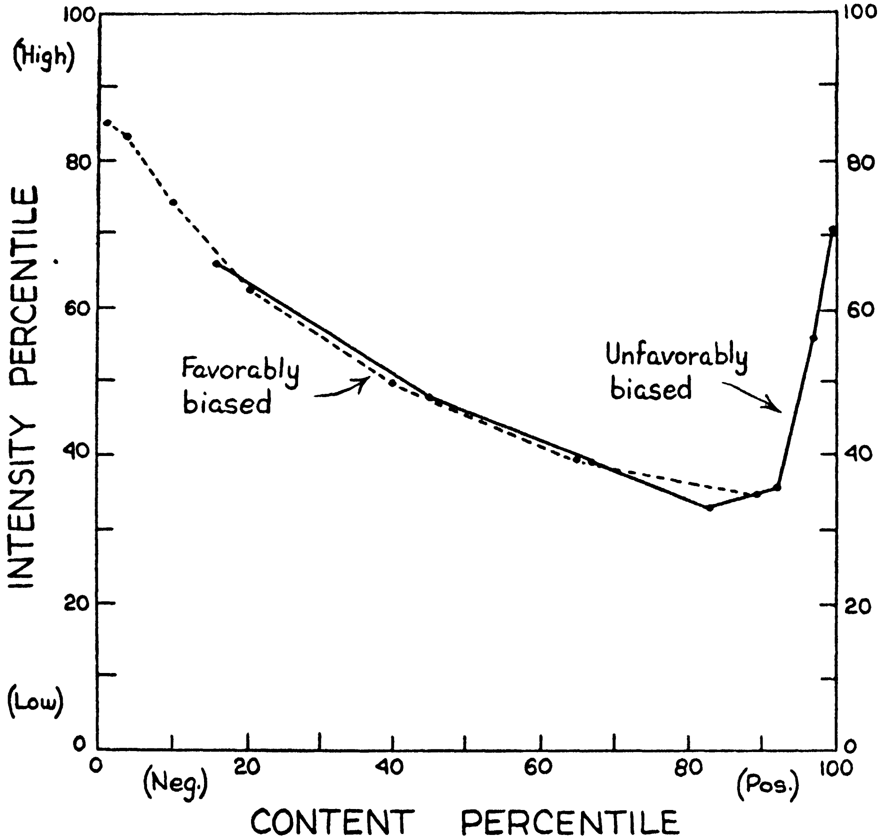


Figure 24. Attitude toward officers.
Series A—positive end of universe (Table 24a—"unfavorably biased").
Series B—negative end of universe (Table 24b—"favorably biased").

would come to the conclusion that about 30 per cent of the men were favorable, and about 70 per cent were unfavorable toward the Army.

The same results are obtained for Series A and B on attitude toward officers. The intensity curve of Series A, with "unfavorably biased" dichotomizations, shows quite clearly that approximately 83 per cent have "unfavorable" opinions of their officers. Despite the fact that Series B shows no more than 50 per cent "unfavorable"

on any of the six questions, its intensity curve would produce an estimate of between 65 and 89 per cent "unfavorable" (midpoint 77 per cent), with a clear-cut indication that additional questions showing greater "unfavorableness" toward officers are needed for a closer determination of the dividing point.

Of course, the best answer would be obtained by combining Series A and Series B and asking the respondent all questions (or selecting alternate questions from each series). The clear-cut results obtained by asking all questions have already been shown for these two areas in the previous section. These previous figures show a careful sampling of all degrees of "favorable" or "unfavorable" opinions, and clearly indicate just how the entire population should be divided, so as to obtain an objective, invariant distribution into "pro" and "con" which is free from the problem of question "bias."

Summary

1. One purpose of intensity analysis is to provide an objective method of determining an invariant cutting point for an attitude or opinion scale. This cutting point will enable the research worker to divide his population into favorable and unfavorable groups, a division which will be independent of the selection or "bias" of the specific questions asked.

2. The basis of intensity analysis is the theory of scaling which provides a test for the single meaning of a series of questions. Any single question or series of questions on an issue is considered as simply a sample of all other questions on the same issue which might have been asked instead. The problem of scaling is to test whether the particular sample of questions used can be considered as belonging to an attitude or opinion universe that concerns only one topic. Once a selected series of questions has been determined to be scalable, it is possible to rank the respondents from high to low on the single attitude universe being studied.

3. In addition to the rank order of individuals on the content or attitude scale, it is also possible to determine the intensity with which an attitude is held. This intensity measurement is found to be a *U*- or *J*-shaped function of the content scale. People on both ends of the content scale feel more strongly than people in the middle of the scale. As one moves down the content scale, intensity of feeling decreases *until a point is reached* where intensity of feeling begins to increase again. This point is invariant for any single

attitude area, and regardless of the sample of attitude questions used will always divide the population into the same *proportion* with positive and negative opinions. Thus, this cutting point is both invariant and objective.

4. Not only is the zero point invariant with respect to item selection, but so also is the entire curve. The shape of the curve around the zero point shows whether there is a broad region of relative indifference or neutrality, or whether the population is sharply divided on the issue.

5. One method used in this chapter for measuring intensity of feeling was to ask, "How strongly do you feel about this?" after each attitude question. Correlating the content scale scores with the intensity scores produced the hypothesized *U*- or *J*-shaped curve. This technique of measuring intensity is far from perfect and results in a large amount of variance around the curve. Further research will probably serve to reduce much of this variance. However, crude as it is, the present technique does work and has been used successfully in several instances. The fold-over technique was also found to succeed occasionally; this technique has the advantage of not lengthening the questionnaire.

CHAPTER 8

PROBLEMS OF RELIABILITY[1]

Three kinds of reliability. There are at least three important sources for variation in empirical results that are to be considered in an attitude or opinion survey. One is the sampling of people; the second is the sampling of items or questions; and the the third is the sampling of trials.

The problem of sampling of people is the usual one which arises when the whole population is not observed, but only a sample therefrom is taken. How much do the sample results vary from what would have been obtained if the whole population of people had been studied? For scalogram analysis, we want to know if the hypothesis that the data are scalable for the entire population can be tested by means of only a sample of the people. In practice, pretests on about a hundred people have been used to decide whether or not an area of content was scalable. To what extent is this justified?

The second problem of sampling is with reference to the universe of items. Any questions used in an attitude or opinion survey are but a sample of all possible similar questions that could have been used. The hypothesis that scalogram analysis intends to test is that the entire universe of items is scalable. In practice, a sample of ten or twelve items in a pretest has been used as a basis for the test. Can such a sample properly test the hypothesis?

The third problem of sampling concerns the stability of responses of individuals to the items. If the survey were to be repeated over and over again on the same individuals and on the same items under similar conditions—without carryover from trial to trial—how much variation would there be in the responses of the individuals? This is the problem of test-retest reliability.

Let us discuss each of these three kinds of reliability problems in turn, from the point of view of scalogram analysis.

With respect to the sampling of people, we shall consider the relia-

[1] By Louis Guttman.

bility of the coefficient of reproducibility and of the frequencies of the various scale types. The coefficient, we shall find, can have high reliability in the small sample of people used for the pretest, while the frequency distribution of types (for the given sample items) requires the larger sample of the final study for stability.

With respect to the sampling of items, we shall consider the reliability of the coefficient of reproducibility and show the need for the additional criterion of "improvement." We shall also study the reliability of the scale scores. The hypothesis of approximately perfect reproducibility will be found to be testable on the basis of a dozen or so items in a pretest. If the hypothesis is accepted, then fewer items can be used in the final study.

As for test-retest reliability, or the sampling of trials, we shall find that approximately perfect scales necessarily have high reliability, both with respect to the individual's response to each item and to his total score.

The Sampling of People

The hypothesis of approximate perfection. In order to develop a sampling theory for any problem, certain characteristics of the population distribution must be specified and the sampling scheme must be defined. The hypothesis to be tested by scalogram analysis is that the population has an approximately scalable distribution of responses for the given items, that is, that each item is highly reproducible from scale scores. Such a hypothesis of approximate perfection differs somewhat from customary hypotheses found in sampling theory elsewhere. The usual kind of hypothesis is that complete *im*perfection exists: that there is *no* difference between groups or that variables are *not* correlated, etc., etc.

It is quite easy to refute a hypothesis about perfection by means of a sample drawn by almost any scheme, random or not. For example, it is easy to prove that height and weight are *not* perfectly correlated for a given population by taking three people and finding that there is not a perfect linear relation between their heights and weights. Three people are sufficient here because if there were a perfect relationship in the population, this would have to hold in any sample whatsoever. If all points lie on a straight line, then any sample of those points must lie on the same straight line. There is no room for sampling variation.

Similarly, if a set of items were perfectly scalable for a population, it would have to be perfectly scalable for any sample of people

whatsoever drawn from that population. If an item is a simple function of the scale scores for the population, it must remain the same simple function of the scale scores in any sample drawn from that population. This is a form of perfect correlation, and there is no room for sampling variability. If, in the population, all people with the same scale score have the same response to an item, then any sample of people with that scale score will have only the given item response.

It is no trick, then, to destroy the hypothesis that a set of items is perfectly scalable. Sampling error with respect to people can occur only when there is imperfection or scale error *in the population.* In practice, we do not find perfect scales in samples, which therefore means that our data are not perfectly scalable for the populations from which the samples are drawn. This leads to difficulties in formulating an analytical statement of the population distribution of responses, because there are many varieties of possible population patterns of error. The coefficient of reproducibility is but one aspect of error, and even its sampling distribution thus far has not yet yielded to a complete mathematical analysis. The sampling theory of scale error is still in its infancy as far as fully rigorous treatment goes.

The reliability of the coefficient of reproducibility. There is, however, cogent reason for believing that the higher the coefficient of reproducibility for the population, the less likely is the reproducibility of any random sample of people to depart from it by any given amount. There is reason to suspect that the sampling behavior of the coefficient should not be radically different from the sampling behavior of ordinary proportions (which follow the binomial distribution) especially when the population reproducibility is high (say, over .90). As is well known for the random sampling of ordinary proportions, if P is the population proportion and if N is the size of each random sample, then the standard error of sample proportions is $\sqrt{\frac{P(1-P)}{N}}$. This standard error decreases as P approaches unity, and in fact is zero if $P = 1$.

It is highly plausible that a similar situation should hold for the coefficient of reproducibility. We do know that the standard error of sample reproducibility coefficients is zero if the population reproducibility is unity, for there is no room for sampling variation in the perfect scale. If there is but little error in the population, then there is little chance for error in a sample, and the more error there

is in the population the more chance for error there is in a sample. Hence, the variance of sample coefficients of reproducibility should increase as the reproducibility of the population decreases. It seems highly plausible, then, that the sampling theory of ordinary proportions can be used as an approximation to the sampling theory of the coefficient of reproducibility. If so, a random sample of a hundred people, with responses to ten items, provides a base of a thousand observations on which to compute the coefficient of reproducibility. If the errors of reproducibility are independent of each other, and if all the items have equal reproducibility for the population, then the sample coefficient of reproducibility would have exactly the same distribution as an ordinary random sample proportion based on a thousand cases. If the items have unequal reproducibility for the population (and errors are independent), then the coefficient would have the distribution of an ordinary proportion of a stratified random sample, which would have a smaller variance than an unstratified random sample with the same average population proportion. The crux of the matter lies in whether or not the errors are independent. An adequate test of this hypothesis of independence has not yet been devised.

In any event, requiring reproducibility of at least .90 tends to insure that, by and large, only populations with reproducibility in that neighborhood will be inferred to be approximately scalable, since it seems clear that the sampling variance of the coefficient of reproducibility cannot be exorbitant in this range.

Empirical studies of variation in reproducibility. Empirical studies of the sampling variability of approximate scale patterns have been made, and they indicate very high stability both for the amount and the general pattern of error. Only one example need be presented in detail here. Seven questions on attitude toward officers were asked on one form (Scalogram 15: Form A) of a questionnaire to a cross section of enlisted men, and a different set of seven questions on another form (Scalogram 15: Form B). The questions are as follows:

Attitude toward officers
Form A

25. Do your officers give you a good chance to ask questions as to the reason why things are done the way they are?
    1 ______ Yes, always
    2 ______ Yes, usually
    3 ______ Undecided

4 ______ No, not very often
5 ______ No, almost never

31. In general, how good would you say your officers are?
1 ______ Very good
2 ______ Fairly good
3 ______ About average
4 ______ Pretty poor
5 ______ Very poor

32. How many of your officers use their rank in ways that seem unnecessary to you?
1 ______ Almost all of them do
2 ______ Most of them do
3 ______ Some of them do
4 ______ Only a few of them do
5 ______ None of them do

34. How many of the officers in your company (battery, squadron, troop) are the kind you would want to serve under in combat?
1 ______ All of them are
2 ______ Most of them are
3 ______ About half of them are
4 ______ Few of them are
5 ______ None of them are

35. How much do you personally like your officers?
1 ______ Very much
2 ______ Pretty much
3 ______ Not so much
4 ______ Not at all

37. How many of your company officers are the kind who are willing to go through anything they ask their men to go through?
1 ______ All of them are
2 ______ Most of them are
3 ______ About half of them are
4 ______ Few of them are
5 ______ None of them are

38. How do you feel about the officers that have been picked for your outfit?
1 ______ They are the best ones that could have been picked
2 ______ They are as good as any that could have been picked
3 ______ Somewhat better ones could have been picked
4 ______ Much better ones could have been picked
5 ______ Undecided

Form B

27. How many of your officers take a personal interest in their men?
1 ______ All of them do
2 ______ Most of them do
3 ______ About half of them do
4 ______ Few of them do
5 ______ None of them do

28. Do you think that your officers generally do what they can to help you?
    1 ______ Yes, all the time
    2 ______ Yes, most of the time
    3 ______ No, they often do not
    4 ______ No, they almost never do

29. How well do you feel that your officers understand your problems and needs?
    1 ______ They are very much aware of my problems and needs
    2 ______ They are fairly well aware of my problems and needs
    3 ______ They do not know very much about my real problems and needs

30. Do you feel that your officers recognize your abilities and what you are able to do?
    1 ______ Yes, I'm sure they do
    2 ______ Yes, I think they do, but I'm not sure
    3 ______ No, I don't think they do
    4 ______ Undecided

33. When you do a particularly good job do you usually get recognition or praise for it from your officers?
    1 ______ Always
    2 ______ Usually
    3 ______ Rarely
    4 ______ Never

36. When your officers give you something to do, do they tell you enough about it so that you can do a good job?
    1 ______ Always tell me enough
    2 ______ Usually tell me enough
    3 ______ Often do not tell me enough
    4 ______ Almost never tell me enough
    5 ______ Undecided

39. Can you count on your officers to back you up in your duties?
    1 ______ Yes, always
    2 ______ Yes, usually
    3 ______ No, you can't count on it

A random sample of a hundred men was drawn for the first form and a scalogram analysis made. The resulting scalogram for Form A is shown in Scalogram 15, A. The coefficient of reproducibility is .89.

The scalogram for Form B, based on the responses of the same men used for Form A, shows a closely similar pattern of error with the coefficient of reproducibility of .92, as shown in Scalogram 15, B.

Similar results have been obtained in many other experiments. Perhaps one more example might be mentioned briefly. In a study of desire to return to school after the war, the pattern of error for a

sample of three thousand enlisted men was studied and found to agree closely with the pattern of error previously obtained in a sample of a hundred men drawn at random from the three thousand.

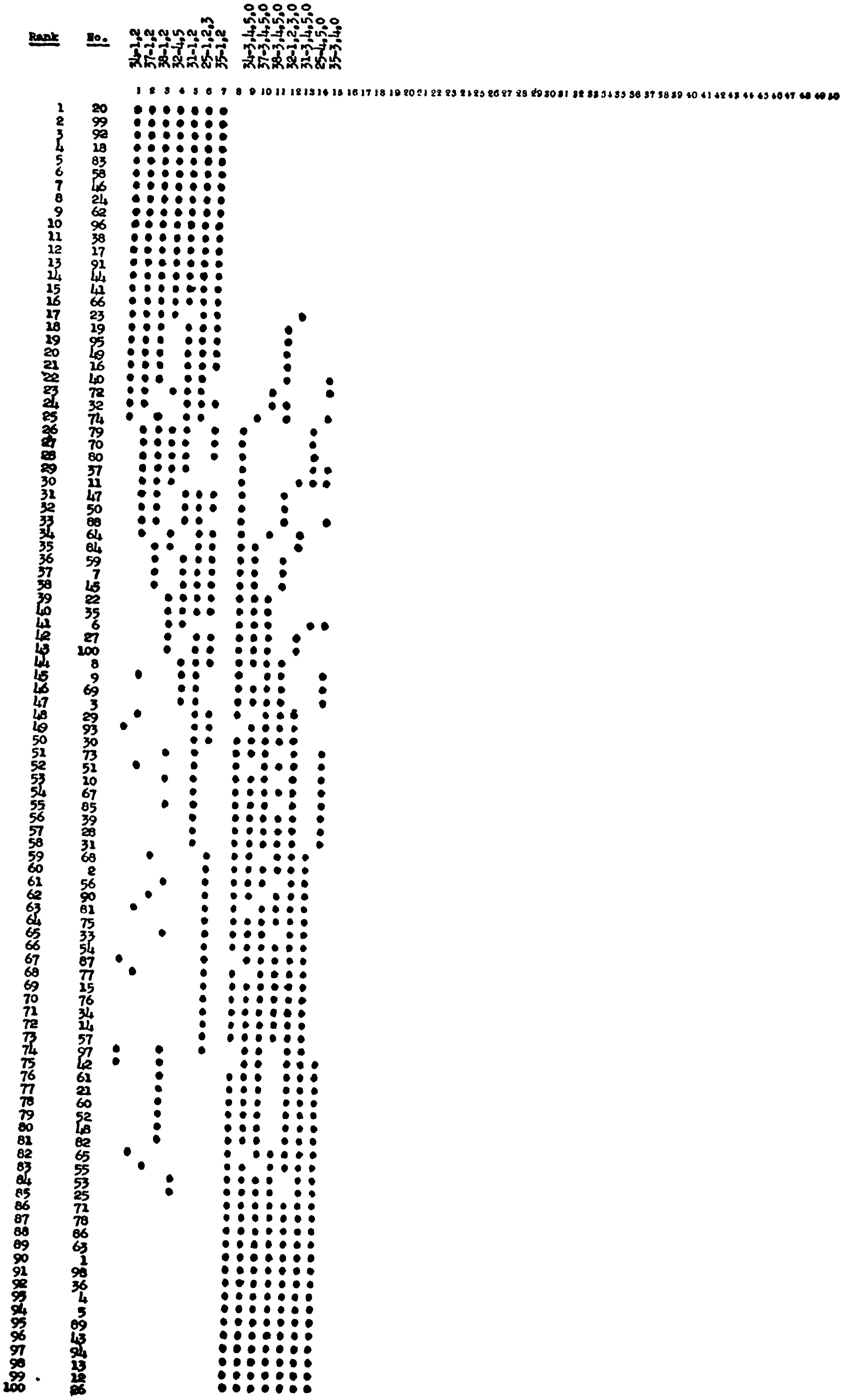
Reliability of marginal frequencies and other aspects. It should not be inferred from the fact that a hundred cases may be adequate to test the hypothesis of scalability that they are therefore adequate for use in the final study to portray the distribution of the attitude or opinion. That is quite a different matter. The proportion of people at each scale type or score is subject essentially to the laws of ordinary random sampling. If the proportion in the population of a given scale type is not close to zero or unity, then the sample proportion can have considerable sampling variation. Such proportions have a base of only a hundred observations, which is a small sample on which to base an ordinary porportion. For example, compare the proportions of each scale type for the two examples of men on attitude toward officers just shown above. The frequency distributions of the scale types differ according as to what might be expected from only a hundred cases. In order to obtain a stable distribution of scale types—which is important for external prediction purposes and for internal purposes like obtaining a zero point by intensity analysis—the usual three thousand or so cases are of the magnitude desired.

A small sample of about one hundred cases may be used in a pretest to test the hypothesis of scalability, and the hypothesis may be rejected or accepted on the basis of the sample. If the hypothesis is accepted, then the final study can be conducted on a larger sample to obtain reliable frequencies for the scale types. Fewer items can be used in the final study than in the pretest, as is discussed next in connection with the problem of sampling of items.

If items with four or five categories are used initially, and if scalogram analysis shows that much combination of categories is required to obtain high reproducibility, then it is desirable to use more than a hundred cases in the pretest. The initial reproducibility is so far from perfect that it may allow for substantial sampling error in the combinations. Similarly, quasi scales can be expected to have a greater sampling variance in their error pattern because of their imperfect reproducibility, and hence require larger samples than a hundred in a pretest. The future undoubtedly will bring sharper mathematical and empirical evidence to bear on the sample size to use for these various aspects of scalogram analysis.

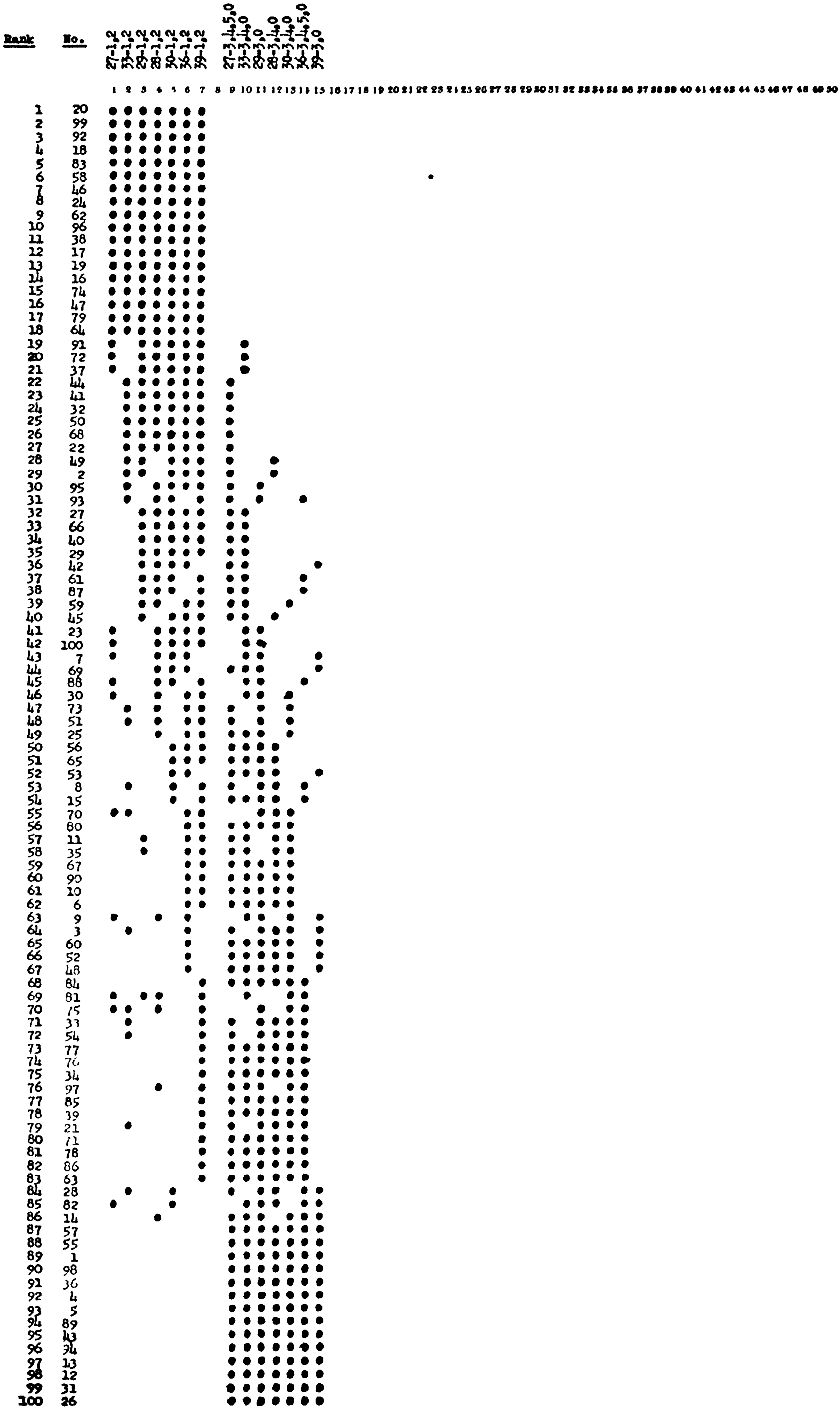
SCALOGRAM No. 15. ATTITUDE TOWARD OFFICERS: FORM A (Coefficient of Reproducibility = .89)

Respondent order *Category order*



SCALOGRAM No. 15. ATTITUDE TOWARD OFFICERS: FORM B (Coefficient of Reproducibility = .92)

Respondent order *Category order*



The Sampling of Items

The hypothesis of approximate perfection. In any attitude or opinion study, only a sample of items is used from the entire universe of content under consideration. It is difficult to think of items as being drawn by anything like a random process from the universe. Questions are *constructed* by the research worker. He selects a particular wording of the question, a particular aspect of the content to emphasize, etc., etc. It is not as if there were available a list of all possible questions and their variations from which those used in the study were drawn at random. It seems that random sampling theory is not adequate for the problem of inference for the case of sampling from a universe of content.

One of the fortunate aspects of scale analysis is that the hypothesis it wants to test differs in its nature from the ordinary hypotheses of statistical inference in other fields. The hypothesis is that an approximately *perfect* pattern exists, whereas most of current statistical inference, as mentioned in the previous section, is concerned with testing hypotheses of *lack* of relationships. Just as the hypothesis of a perfect linear correlation can be destroyed by a sample of three people, whether drawn at random or not, just so can the hypothesis of a perfect scale be destroyed by a sample of three—or even two—items, no matter how they are drawn. Therefore, to test the hypothesis of a perfect scale we do not have to worry about the fact that items are not necessarily selected at random from the universe of content. No matter how they are selected, they can serve to refute the hypothesis that the universe of content is perfectly scalable for the population.

In practice, the hypothesis of perfect scalability has always been refuted by the samples obtained. For example, we never find in a fourfold table between two items that there is a perfectly zero cell as required by the perfect scalogram pattern. (See Chapter 3.) This means that neither item can be perfectly reproducible from scale scores derived from the universe of items, for they are not even perfectly reproducible from scores derived only from themselves.

In practice, we are concerned not with a hypothesis of perfect reproducibility, but rather a hypothesis of *approximately* perfect reproducibility. Can we infer, if the reproducibility of a sample of items is .95, that therefore the reproducibility of the entire universe of items is in a relatively small neighborhood of .95? If a million items have a reproducibility of .90, how much variation would there

be in the reproducibility of all samples of ten items each that could be drawn from this larger set? These questions do not seem susceptible to solution by ordinary random sampling theory because, as was indicated just previously, items are not selected by any random process. Furthermore, we have no way yet of specifying the population distribution of items.[2] How many items are there in the universe which will split the population into two groups of 50–50? How many will split the population into a 60–40 division? How many into a 70–30 division? Etc., etc.

Reproducibility and improvement as criteria. In order to avoid having to use a notion of random sampling, it seems that we must use a criterion which will involve again some notion of perfection so that *there will not be much room for variation regardless of how the sample is drawn.* Then it will be relatively immaterial whether the sample is random or not.

In practice we have used two criteria. One is that each item separately should have high reproducibility. This is a kind of perfection. If *all* items in the universe have reproducibility of at least .85, then every item drawn from the universe, whether random or not, will have reproducibility of at least .85. Hence, if in a sample of ten items, each item separately has reproducibility of at least .85, it would seem plausible to infer that the universe reproducibility was at least that high. Therefore, a criterion used is not only that the total sample reproducibility shall be around .90, but that each separate item should have reproducibility not much below .90.

A second criterion is to take care of the possibility that sample reproducibility is spuriously high. It is easy to see that for any item whatsoever, whether it belongs to a scalable universe or not, the reproducibility cannot be less than its highest category frequency. For example, in a dichotomous item where 90 per cent of the population is in one category and 10 per cent is in the other, there cannot be more than 10 per cent error for reproducing that item from any set of scores whatsoever, regardless of the relationship of that item to the other items or to the scores. By simply predicting every person to be in the category with the largest frequency, the proportion of correct estimates will be the relative fre-

[2] The population distribution of items determines the shape of the intensity and other components of a scale, according to the equations in the next chapter. Hence, by obtaining an empirical intensity function as in Chapter 7, we obtain indirect information about the universe of items. It seems plausible that studying the further components empirically may be a proper approach to determining the universe distribution of items. This would avoid the problem of obtaining random samples of items.

quency of that modal category. Hence, it is important to guard against spuriously high sample reproducibility for items which have modal categories which contain a vast majority of the population. To do this, we use the criterion of *improvement*. Not only must reproducibility of each item be high from the trial scale score, but the scale error must be at most half of that which would be obtained without knowledge of the scale pattern, that is, from the modal frequencies alone.

An example of rejecting the scale hypothesis. Let us illustrate how high reproducibility can be obtained from a sample of items from a nonscalable universe, and yet the hypothesis of scalability will be rejected by means of the criterion of improvement. In order to keep the data simple, we shall consider a case of four dichotomous items which are statistically independent of each other. If these are "Yes"-"No" questions, by statistical independence is meant that the proportion of people who say "Yes" simultaneously to two questions is equal to the product of the proportion of those who say "Yes" to the first question and the proportion who say "Yes" to the second question; the proportion who say "Yes" to the first question but "No" to the second question is equal to the product of the proportion of those who say "Yes" to the first and the proportion who say "No" to the second question; and similarly for the other two combinations of responses. Statistical independence is the condition of *no* relationship whatsoever between the items, an opposite of scalability, which means that each item is a perfect function of the same variable.

If the four items are statistically independent of each other, then we can calculate not only the joint occurrences two at a time of the items from their separate frequencies, but also the higher order joint occurrences, including four at a time. For example, the proportion of the population who say "Yes" to all four questions is equal to the proportion who say "Yes" to the first question *times* the proportion who say "Yes" to the second question *times* the proportion who say "Yes" to the third question *times* the proportion who say "Yes" to the fourth question. There are sixteen possible combinations of responses from four dichotomous questions, and the proportion of people who will fall into each group can be calculated by forming products in this manner if we know the frequencies of Yes's and No's for each item separately.

Let us assume that the first question has 20 per cent "Yes" and 80 per cent "No," the second has 40 per cent "Yes" and 60 per cent

"No," the third question 60 per cent "Yes" and 40 per cent "No," and the fourth question 80 per cent "Yes" and 20 per cent "No." Given further that the four items are statistically independent (and hence nonscalable), we can calculate the frequency for each of the sixteen possible kinds of combinations of responses to the four questions, perform a scalogram analysis on these, and wind up with the

TABLE 1

FOUR INDEPENDENT DICHOTOMIES

Group	*Scale type*	*Category*								*Frequency*
		Yes$_1$	*Yes*$_2$	*Yes*$_3$	*Yes*$_4$	*No*$_1$	*No*$_2$	*No*$_3$	*No*$_4$	
1	4	x	x	x	x					.0384
2	4	x	x	x					x	.0096
3	4	x	x		x			x		.0256
4	3		x	x	x	x				.1536
5	3		x	x		x			x	.0384
6	3		x		x	x		x		.1024
7	2			x	x	x	x			.2304
8	2	x		x	x		x			.0576
9	2			x		x	x		x	.0576
10	2	x		x			x		x	.0144
11	1				x	x	x	x		.1536
12	1	x			x		x	x		.0384
13	0					x	x	x	x	.0384
14	0	x					x	x	x	.0096
15	0	x	x					x	x	.0064
16	0		x			x		x	x	.0256
Marginal frequency		.20	.40	.60	.80	.80	.60	.40	.20	1.000
Error		.1264	.0320	—	—	—	—	.1280	.1200	
Nonerror		.0736	.3680	.6000	.8000	.8000	.6000	.2720	.0800	

scalogram arrangement shown in Table 1. The scalogram picture in Figure 1 will show more clearly the distribution of error. In Table 1, the x's indicate the responses, but the frequency is shown in the right-hand column, since we have shown but one x for each group instead of repeating enough x's to represent the proportion of people in the group. Figure 1 shows the frequencies of each type graphically, so that error can be estimated visually. Continuous line segments are used in Figure 1 instead of the usual columns of dots; the length of each segment shows the per cent of cases in the

category at the given rank intervals. Scale types are defined as in Chapter 4, according to the criterion of least number of errors.

The reproducibility of this sample of four items is .90. However, we would not infer therefore that the universe from which these four items are drawn has a reproducibility of .90; the second criterion of improvement is not satisfied. At least two of the items have spuriously high reproducibility according to our second criterion. The first and fourth questions, because of their marginal frequencies,

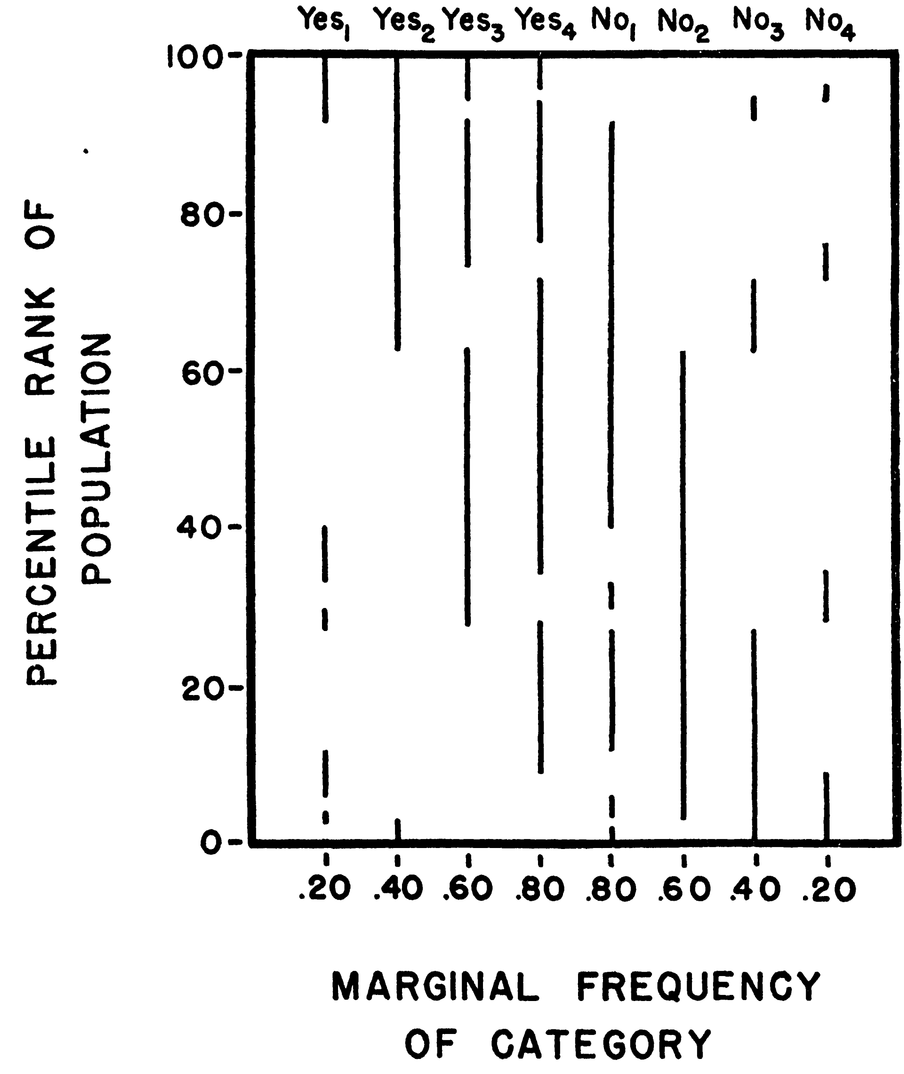


Figure 1. Four independent dichotomies.

cannot have more than 20 per cent error regardless of scalability because each has 80 per cent of the cases in its modal category. The reproducibility of these items in the sample is higher than 80 per cent because the scale arrangement is based partly on the items, so that there is some improvement over 80 per cent since the item itself is used in a partially circular fashion. Actually, the first item has 13 per cent error in its first category, which is more error than nonerror for this nonmodal category. Similarly, the fourth question has 12 per cent error in its "No" category, which is also more than half the frequency of its nonmodal category. Neither of these items has had its error cut down by at least half, so we conclude that their reproducibility is spuriously high.

Since two of the four questions, or half the sample of items, are judged to have spuriously high reproducibility, we reject the hypothesis that the universe from which they are drawn is scalable.

It is clear that if we had used more than four independent items, we would have been all the more certain to reject the hypothesis of scalability. The more items in the sample, the more possible is it for error to appear if it is present in the universe of items.

An empirical example. If the items are not independent but have positive intercorrelations, they still need not be scalable. Using about a dozen items has been found empirically to be rather satisfactory in distinguishing between cases which are sufficiently scalable and those which are not. As an example, the following mixture of thirteen items was subjected to a scalogram analysis:

27. How satisfied are you about being in your present Army job instead of some other Army job? (Check one)
    ______ Very satisfied
    ______ Satisfied
    ______ It does not make any difference to me
    ______ Dissatisfied
    ______ Very dissatisfied

34. After the war do you think you would like to come back to your old outfit for two weeks of training every year? (Check one)
    ______ Yes, I would want to very much
    ______ Yes, I might want to, but I'm not sure
    ______ No, I don't think I would like to
    ______ No, I would not want to at all
    ______ Undecided

36. If you were sent into actual fighting right now, how do you think you would do? (Check one)
    ______ I think I would do all right
    ______ I think I would have trouble at first, but after a while I would be O.K.

______ I don't think I would do very well
______ I haven't any idea how I would do

45. Do you think that after the war soldiers in your outfit will find it easier or harder to get jobs than they did before the war? (Check one)
______ Easier than before the war
______ About the same as before the war
______ Harder than before the war
______ Undecided

57. Do you think it is at all likely that we will fight England in the next 25 years or so? (Check one)
______ Not possible
______ Not very likely
______ Quite likely
______ Almost certain

60. How much do you think the United States is doing to win the war? (Check one)
______ Doing all it possibly can
______ Doing all it can in most things
______ Could do quite a bit more
______ Could do a whole lot more

64. How do you think this war will come out? (Check one)
______ It will be a complete victory for us
______ We will win, but it won't be a complete victory
______ It will end in a draw
______ The other side has a pretty good chance of winning
______ Undecided

67. Do the men in your company (battery, squadron, troop) co-operate and work well together? (Check one)
______ All of the time
______ Most of the time
______ Often do not
______ Almost never do
______ Undecided

75. Do your noncoms give you a good chance to ask questions as to the reason why things are done the way they are? (Check one)
______ Always
______ Usually
______ Not very often
______ Almost never
______ Undecided

79. How many of your company officers are the kind who are willing to go through anything they ask their men to go through? (Check one)
______ All of them are
______ Most of them are
______ About half of them are
______ Few of them are
______ None of them are

89. Do you think the equipment of the American Army is better or worse than the equipment of the German Army? (Check one)
______ All of our equipment is better
______ Most of our equipment is better
______ Our equipment is about the same as that of the German Army
______ Most of our equipment is worse
______ All of our equipment is worse
______ Don't know

92. In general, how serious an offense do you think it is for a solider to go AWOL? (Check one)
______ Very serious
______ Pretty serious
______ Not so serious
______ Not serious at all
______ Undecided
______ It depends on the condition

96. Up to now, has the war made things better or worse for your family and friends back home (such as jobs, pay, and living conditions)? (Check one)
______ Things are a lot better
______ Things are a little better, but not much
______ I don't see any difference
______ Things are a little worse
______ Things are a lot worse
______ Don't know

These questions are not drawn from any one universe of content and are not expected to be scalable, although it is known there are positive correlations between the various kinds of content involved in this mixture. The reproducibility is only .82. Furthermore, practically every question would be suspect on the basis of the criterion of improvement. This result is obtained here even though the questions were not originally dichotomies so that there was more room for sampling variation and for the appearance of spuriously high reproducibility through the combination of categories.

An example of trichotomies. The hypothesis of scalability seems even more tenable if high reproducibility and proper improvement are obtained when three or more categories can be retained in at least some of the items. For example, on the basis of only three dichotomies it will usually be hard to reject the hypothesis of scalability even though it is false, but it is easier to reject the hypothesis if the items are trichotomous. In Table 2 and Figure 2 are shown a scalogram analysis of three independent trichotomies. The items are indicated by the numbers 1, 2, 3, and the categories by the letters a, b, c. The twenty-seven possible combinations of answers

have their relative frequencies computed, assuming that the three items have the marginals indicated and are independent. The reproducibility is .84, and each of the three items has a category with more error than nonerror. It is clear, then, that *if items can remain in trichotomous form in a scalogram analysis, it is more plausible that the universe is scalable* than if they have to be combined. It is easier for error to appear—if it actually should be present—when

TABLE 2

THREE INDEPENDENT TRICHOTOMIES

Group	*Scale type*	*Category*									*Freq.*
		1a	*2a*	*3a*	*1b*	*2b*	*3b*	*1c*	*2c*	*3c*	
1	6	x	x	x							.006
2	6	x	x				x				.012
3	6	x	x							x	.002
4	6	x		x					x		.006
5	5		x	x	x						.036
6	5		x		x					x	.012
7	5		x	x				x			.018
8	4			x	x	x					.108
9	4	x		x		x					.018
10	4			x	x				x		.036
11	3				x	x	x				.216
12	3				x	x				x	.036
13	3	x				x	x				.036
14	3		x		x		x				.072
15	3				x		x		x		.072
16	2	x				x				x	.006
17	2					x	x	x			.108
18	2		x				x	x			.036
19	2			x		x		x			.054
20	2					x		x		x	.018
21	1	x					x		x		.012
22	1						x	x	x		.036
23	0			x				x	x		.018
24	0							x	x	x	.006
25	0		x					x		x	.006
26	0	x							x	x	.002
27	0				x				x	x	.012
Marginal frequency		.10	.20	.30	.60	.60	.60	.30	.20	.10	
Error		.074	.114	.072	.012	.000	.012	.018	.114	.074	
Nonerror		.026	.086	.228	.588	.600	.588	.282	.086	.026	

more categories are used; if error does not appear in this case, it is more plausible to assume that there was not much error really to be expected in the population.

The reliability of scale ranks. Thus far we have discussed the sampling of items with respect to reproducibility. A related aspect is the stability of the ranking of pèople. To what extent does the rank order of people on a sample of items correspond to the rank order that would be obtained on the universe of items?

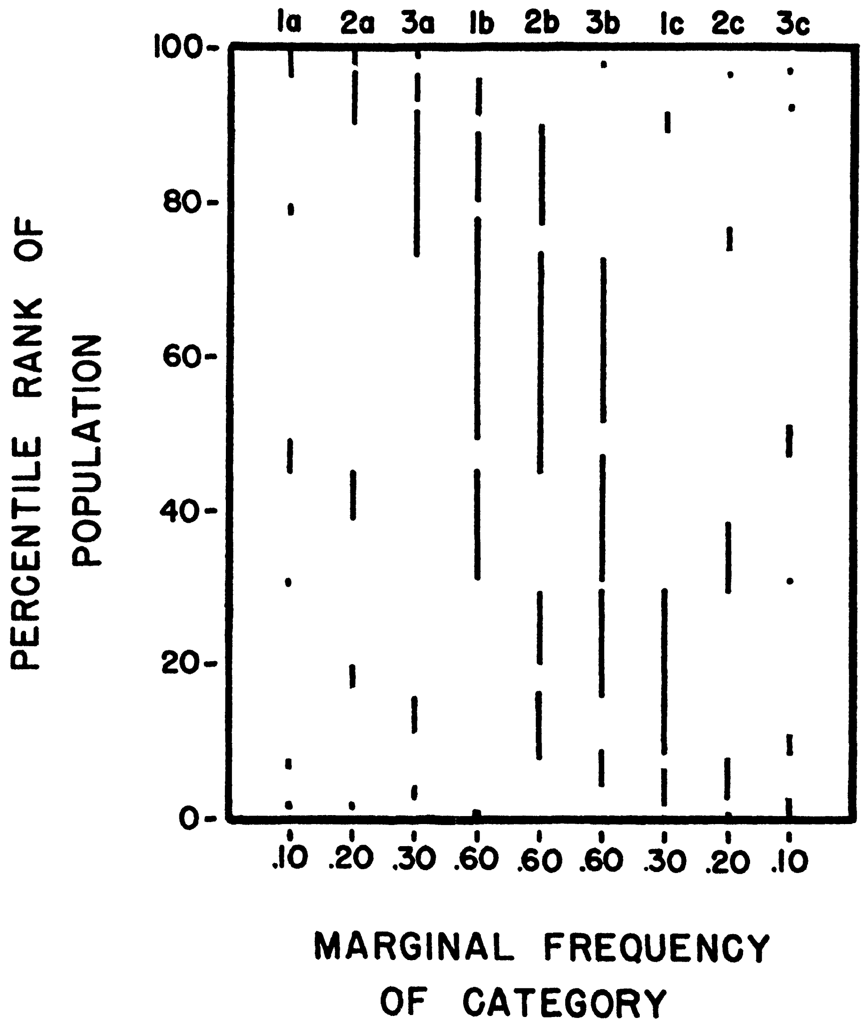


Figure 2. Three independent trichotomies.

In one sense, the correspondence can never be perfect between sample and universe ranks even for the case of a perfect scale. For example, a sample of ten dichotomous items from a perfect scale can yield at most eleven distinct ranks. In the universe of items, there may be an infinite number of distinct ranks. The sample ranks, therefore, represent groupings or class intervals of the universe ranks, and there would not be a one-to-one correspondence between sample ranks and universe ranks. The sample ranks will be a simple function of the universe ranks, but not vice versa. Two people can have the same sample rank but differ somewhat in their universe rank. Moreover, two people with different sample ranks must also have different universe ranks. One person who is higher than another in the sample must be higher than the other in the universe, if the universe is perfectly scalable, whereas two people with the same sample rank may be either in the same or in adjacent ranks in the universe.

In practice, even for a perfectly scalable universe, it would not usually be desirable to work with a large number of ranks. In much of attitude and public opinion research it is sufficient to work with deciles or quintiles or even grosser classifications. The infinite number of ranks in the universe, therefore, is not needed; the ranks obtainable from a sample can be quite sufficient for the purpose of the research. If finer rankings are desired like percentiles, then of course a larger sample of questions is needed. For the perfectly scalable universe, then, the number of questions to use in a sample depends upon the number of ranks between which it is desired to differentiate for the given research problem.

For the case of an imperfectly scalable universe, which is the usual case in practice, we have to worry about not only the ultimate number of ranks desired for a particular research project, but also the reliability with which an individual is placed into his proper rank on the basis of the sample. The unreliability we are concerned with here is that due to the sampling of items. Just like the coefficient of reproducibility, there can be no sampling variation in the fact that the sample ranks must be a simple function of the universe ranks for the case of a perfectly scalable universe, and the possibility for unreliability in sample ranks increases as the reproducibility of the universe goes down. Requiring that reproducibility be about .90 or higher means that as a consequence we obtain very substantial reliability for the ranking of individuals.

Relationship between part score and whole score. As an empirical

example, let us consider the area of attitude toward officers. A sample of eleven questions was used in this area, each of which was dichotomized in the course of the scalogram analysis. Thus, total scores in the sample of eleven questions ranged from 0 to 11. The sample was broken into two parts, one of five questions and the other of six questions, by arranging the questions in order of the frequency of "favorable" responses and selecting every other question. The subsample of five questions we shall call the "first half," and the subsample of six questions we shall call the "second half" of the sample.

The final survey for these items was made on a cross section of three thousand enlisted men. The correlation between the scores

TABLE 3

THE CORRELATION BETWEEN HALF AND WHOLE SCORES

Whole score	*Score on first half* 0	1	2	3	4	5	*Total*
11						.01	.01
10					.01	.03	.04
9					.04	.02	.06
8				.02	.06	.02	.10
7				.05	.05		.10
6			.02	.08	.02		.12
5			.06	.06			.12
4		.02	.07	.03			.12
3		.06	.04				.10
2	.01	.08	.02				.11
1	.03	.05					.08
0	.04						.04
Total	.08	.21	.21	.24	.18	.08	1.00

on the first half and the total scores is shown in Table 3. Table 3, of course, does not show what we are actually interested in. What we would like to know is what the relationship is between scores on a sample of questions and scores on the entire universe. Since we cannot study an entire universe in practice, what we are doing here is showing relationship between the scores on a subsample of questions and scores on the total sample.

The relationship in Table 3, of course, may be somewhat spuriously high because the subsample is included in the total score. On the other hand, it should be remembered that for discrete data with limited possible intervals such as we have in this sample, there is a maximum possible correlation for given marginal frequencies. Given

that the scores on the first half and that the total scores have the respective marginal frequency distributions given in Table 3, the best possible correlation table that they could have is shown in Table 4. Table 4 is what the joint distribution of these two scores would be if both were simple functions of the same quantitative variable. It is impossible for the part score and total score to be perfectly related because their marginal frequencies are so different. At best, both the scores can be perfect functions of a third score which, in this case, would presumably be the rank on the universe of items.

If we take the scale ranks assigned to the first half and to the total at their face value as actual numbers, and if we use the ordinary product-moment coefficient of correlation, we find the correlation between the first half and the total score in Table 3 to be .93, and the maximum possible correlation in Table 4 to be .97.

TABLE 4

MAXIMUM POSSIBLE CORRELATION OF FIRST HALF WITH WHOLE SCORE FOR FIXED MARGINAL DISTRIBUTIONS

Whole score	*Score on first half*						*Total*
	0	*1*	*2*	*3*	*4*	*5*	
11						.01	.01
10						.04	.04
9					.03	.03	.06
8					.10		.10
7				.05	.05		.10
6				.12			.12
5			.05	.07			.12
4			.12				.12
3		.06	.04				.10
2		.11					.11
1	.04	.04					.08
0	.04						.04
Total	.08	.21	.21	.24	.18	.08	1.00

Relationship between two half scores. Another way of looking at the matter of reliability of ranks which will eliminate the possible element of spuriousness in Table 3 is simply to correlate scores on the first half with scores on the second half. This does eliminate the possibility of spuriousness but, on the other hand, is a bit more removed from the direct problem of finding out the correspondence between the sample rank and the universe rank. What we shall be doing here is to investigate the correspondence between rank on one sample and rank on another sample. The table for the correlation

between scores on the first and second halves of the sample is given in Table 5.

Again, since the marginal distributions of the two half samples are different, their scores cannot possibly have a perfect correlation.

TABLE 5

CORRELATION BETWEEN SCORES ON THE TWO HALF SAMPLES

Score on second half	*Score on first half* 0	1	2	3	4	5	*Total*
6					.01	.01	.02
5				.02	.04	.03	.09
4			.02	.05	.06	.02	.15
3		.02	.06	.08	.05	.02	.23
2	.01	.06	.07	.06	.02		.22
1	.03	.08	.04	.03			.18
0	.04	.05	.02				.11
Total	.08	.21	.21	.24	.18	.08	1.00

The maximum possible correlation between the two halves, given the marginals of Table 5, is shown in Table 6. At best, the two half sample scores can be simple functions of a third variable, the universe ranks, in which case their joint occurrence is that shown in Table 6. Treating the half scores as though they were actual numbers and computing the ordinary product-moment coefficient between them yields a correlation coefficient of .74 for Table 5 and .98 for the maximum possible in Table 6.

TABLE 6

MAXIMUM POSSIBLE CORRELATION BETWEEN THE TWO HALF SAMPLES FOR FIXED MARGINAL DISTRIBUTIONS

Score on second half	*Score on first half* 0	1	2	3	4	5	*Total*
6						.02	.02
5					.03	.06	.09
4					.15		.15
3				.23			.23
2			.21	.01			.22
1		.18					.18
0	.08	.03					.11
Total	.08	.21	.21	.24	.18	.08	1.00

Increased reliability with number of items. If the errors of reproducibility behave at all like random errors of measurement, we

should expect that, as the number of items increases, then the relationship will become closer between the ranks obtained from two separate samples. We would expect a much higher relationship between the total scores on the present eleven items and the total scores on another eleven items from the same approximately scalable universe than we find in Table 5 between samples half the size. We would expect this, not because it is necessarily true that reliability of scores (with respect to samples of items) automatically must increase with the number of items, but because we are dealing here with a special case of an approximately scalable universe. It is not true in general for nonscalable universes that adding more items to a sample will increase stability of scores with respect to sampling of items. It is because we are working with a particularly simple kind of universe which has but one dimension, apart from scale error, that the proposition is true about increased reliability with respect to the sampling of items as the number of items increases.

It would not be safe to employ anything like the Spearman-Brown prophecy formula often used by psychologists in this connection, because we know that the assumptions behind such a formula in general cannot be fulfilled by samples even from an approximate scale. Even if there were no scale error, sample ranks will in general not be perfectly correlated with each other. They would be simple functions of the universe scale scores, but that is quite different from being the perfect functions of each other which is required by the Spearman-Brown formula.

How many items to be used in the final study? If a scalogram analysis of a pretest of a dozen or so items shows reproducibility to be about .90 or more, and if the criterion of improvement is also satisfied, then it seems safe to infer that the ranking of people obtained from the sample of items is a good approximation to the ranking that would have been obtained from the universe (with respect to the number of ranks desired). If the items are dichotomies, this means that about a dozen ranks are ascertainable by means of the sample. If there are more than two categories retained in some or all questions, then even more ranks can be determined by the sample.

In many studies it is desirable to use only four or five ranks or even less, since only gross comparisons between groups are intended to be made. For example, in the accompanying volumes on the content of the Research Branch's work, the population is divided very

often only into thirds. In such a case, all the items of the pretest need not be retained for the final study, but only enough items to provide the desired ranks can be used. Since error is present, it is better to use more items than would have been used if no error were present. For example, if it is desired to divide a population into thirds, perhaps five dichotomies might be used to obtain six ranks, and then these ranks can be combined into three ranks with assurance that this will be a fair representation of the ranking in the universe of items. It is doubtful whether even as many as a dozen questions are necessary in the final study for many problems which intend to use the ranks as rough quantitative numbers by treating them as though they were actual integers. In the example just discussed in Table 3, the correlation between scores on five items is .93 with the scores obtained on eleven items. For practical purposes, then, these five items could be used instead of the eleven almost in any situation where it was originally intended to use the eleven.

This is one of the powerful features of the scale pattern. Relatively few items can be used in practice to represent the entire universe of items, with assurance that the predictive power of the universe is being maintained. Adding more items to the sample will not in general enhance the correlations of the scale scores with any outside variable to any appreciable extent.

Reliability of the intensity function. It is the reliability of scale ranks with respect to sampling of items that determines the invariance of the intensity function discussed in Chapter 7. If content scale ranks on one sample of items are closely related to content scale ranks on another sample of items, and if intensity scores on the first sample are also closely related to intensity scores on the second, then the scattergram relating intensity to content cannot be very dissimilar between the two samples. When the data are plotted in the percentile metric, the zero points must approximately coincide, and the entire empirical intensity curves cannot deviate substantially from each other.

Since intensity as obtained by present techniques is a quasi scale, it needs more items in a sample to ensure sufficient stability with respect to the universe of items. Furthermore, since current intensity scores are not pure functions of the content, the shape of the *empirical* intensity curve may not be reliable with respect to *sampling of people* unless a large sample is drawn from the population. If there were no error in the relation between empirical intensity and content, there would be no room for error with respect to the

sampling of people; the median intensity for each content score would have no dispersion. In practice, since there may be considerable spread about the columnar median intensities in the intensity-content scattergram, the sample medians can have correspondingly large sampling error with respect to people unless the large sample of the final study is used. In particular, the rank containing the interval with minimum intensity may not be safely determinable from only a small sample of people.

Again, even if this rank containing the zero interval could be reliably ascertained on the basis of a pretest, the proportions of "positive" and "negative" people in the population would still have the ordinary sampling distribution of a proportion. The usual number of people used in a final opinion survey is needed to determine these proportions reliably.

Intensity, as obtained by current techniques, in general cannot be studied very adequately only in the pretest.

Test-Retest Reliability

Variability within the respondent. The third problem of reliability with which we are concerned might be called that of variability *within the individual respondent.* If he were asked the same question over and over again under the same conditions—without carryover from experiment to experiment—how much variation would there be in his responses? This problem of test-retest reliability is the usual one of errors of observation, but it has often been confused with the problem of sampling of items. It is important to keep these two reliability problems separate and distinct, for each requires a different analysis.

It is only very recently that the theory of the test-retest reliability of qualitative data has been developed.[3] A person's response to a dichotomous question, for example, is to be called completely unreliable, or said to have zero reliability, if in the series of repeated trials he would give one response half the time and the other response the other half of the time. Crudely speaking, it might be said in such a case that the person's response does not differ from what would be expected by "chance." If the person does have a tendency to pick one category over the other, that is, if he chooses one category more than half the time, then he is said to have some reliability, or reliability greater than zero. If he always chooses the

[3] Louis Guttman, "The Test-Retest Reliability of Qualitative Data," *Psychometrika*, Vol. 11, No. 2 (June 1946), pp. 81–95.

same category in all the trials so that there is no variation in his response from trial to trial, he is said to have perfect reliability or reliability of unity. A coefficient to express the various degrees of reliability as numbers between zero and unity is based on the frequency of the category which the individual would choose most often in a series of trials. The category he would choose most often is called his *modal category*, and the relative frequency of this category is called his *modal probability*.

The same concepts hold for items with more than two categories. In the case of a dichotomy, the modal probability for each person is always at least $\frac{1}{2}$, since at least one of the categories must have a probability that big for him. In the case of a trichotomy, the modal probability for each individual is at least $\frac{1}{3}$. The individual will have zero reliability for a trichotomy if he appears equally often in all three categories during the universe of trials. If he is in one of the categories more than $\frac{1}{3}$ of the time, then he will be said to have some reliability. Again, if he is in one of the categories all the time, he will be said to have perfect reliability. Similarly, for an item of four categories the modal probability for each individual is at least $\frac{1}{4}$; the individual will have zero reliability if his modal probability is $\frac{1}{4}$, and his reliability increases as his modal probability exceeds $\frac{1}{4}$.

In general, if an item has m categories, the individual is said to have zero reliability if he is equally often in each of the m categories during the universe of trials; in this case his modal probability is $1/m$. If an individual's modal probability is greater than $1/m$, he will be said to have reliability greater than zero; if he is in one category all the time he will be said to have reliability of unity.

A formula which expresses these various degrees of reliability is as follows. Let P_i be the modal probability for the ith person in the population, and let R_i be his reliability coefficient. The modal probability, P_i, can never be less than $1/m$ and, of course, never greater than unity. The relationship between R_i and P_i is given by the following formula:

$$R_i = \frac{m}{m-1}\left(P_i - \frac{1}{m}\right)$$

If P_i is $\frac{1}{m}$, then R_i is zero. If P_i is unity, then R_i is also unity. R_i departs from zero as P_i departs from $\frac{1}{m}$.

No need to perform a universe of trials. In practice, it is very difficult to perform a series of trials of the same individual on the same question without any carryover from trial to trial. Fortunately, it is not necessary to perform the series of trials in order to learn about the *average* reliability of the responses of the population of individuals. Only a *single* trial is needed in practice to provide information about what would happen if the universe of trials were to be performed. This is possible if a large population of individuals is observed on the single trial.

If we let α be the arithmetic mean of the P_i for the population of individuals, then we can deduce a minimum for α on the basis of only a single trial. Accordingly, if we define ρ to be the arithmetic mean of the reliability coefficients, R_i, we can also state a minimum for ρ on the basis of a single trial. The relationship between the average reliability coefficient, ρ, and the average modal probability, α, is as follows:

$$\rho = \frac{m}{m-1}\left(\alpha - \frac{1}{m}\right)$$

If α is equal to $1/m$, then ρ equals zero; in this case each of the individual P_i is equal to $1/m$ and each of the R_i equals zero. If α equals unity, then ρ also equals unity; in this case each of the P_i is unity and each of the R_i is also unity. An intermediate value of α implies an intermediate value of ρ; this is an average value for the population around which the individual reliabilities cluster.

A single trial will not suffice to estimate α or ρ exactly, but it will suffice to give lower bounds or underestimates of α and ρ. Two kinds of lower bounds have been developed, the second of which is more powerful than the first.[4] Both of these lower bounds are very simple to compute.

The marginal lower bound. The first lower bound can be established from the marginal frequencies of the item itself. The modal proportion of the item *for the population of people* is a *lower bound* for α. For example, in a dichotomy, if 80 per cent of the people are in one category and 20 per cent are in the other, then α is *at least* equal to .80. The average probability in a universe of trials —if it were to be performed—with which an individual would be found in his modal category is at least .80. Using this lower bound for α in the formula relating α to ρ, we obtain a corresponding lower bound to ρ:

[4] *Ibid.*

$$\rho \geqq \frac{2}{1}\left(.80 - \frac{1}{2}\right) = .60$$

The average reliability coefficient for the group thus is *at least* .60. It may be perfect; it may be .90 or anything else down to .60, but it is not less than .60.

Similarly, for an item with four categories, say, if 60 per cent of the people are in one of the categories, then this is the modal category for the population; and the average modal probability for the individuals in the universe of trials is at least .60. The reliability coefficient for the group, then, has the following lower bound:

$$\rho \geqq \frac{4}{3}\left(.60 - \frac{1}{4}\right) = .47$$

Lower bounds computed for the group reliability coefficient this way are called *marginal* lower bounds because they are based on the marginal frequencies of the distribution of the population on the item. In practice the marginal lower bound can grossly underestimate the group reliability coefficient. The second kind of lower bound is more efficient than the marginal lower bound. It is based on the correlation of the item with any other item or variable which is experimentally independent of it. For scale analysis, the best variable to pick is the scale score. For a discussion of the general application of the second kind of lower bound, the reader is referred to the basic article.[5] We shall concern ourselves here only with how to bound the reliability of each item from the scalogram pattern.

Reproducibility as a lower bound. From general considerations of scale theory, it should be clear that if a set of items has high reproducibility, then the items must necessarily have high test-retest reliability. If there were a substantial unreliability factor operating in the responses to the items, this would create appreciable scale error; there would be more than a single factor present. Hence, if scalogram analysis shows that essentially only a single factor is operating in the responses, this must mean that there cannot be many additional factors, including unreliability. This can now be shown explicitly in terms of the second kind of lower bound to the reliability coefficient.

If all items were equally reproducible, then their common reproducibility would be the reproducibility coefficient for the set of items.

[5] *Ibid.*

If, furthermore, each item were experimentally independent of the scale score—that is, if the variation from trial to trial in the universe of trials of each person's response to the item and his scale rank were independent of each other—then the reliability coefficient of the set of items is a lower bound to the average modal probability, α, for each item separately. Therefore, in such a case, if r is the reproducibility coefficient, then a lower bound to the reliability of each item is given by:

$$\rho \geqq \frac{m}{m-1}\left(r - \frac{1}{m}\right)$$

If the reproducibility, r, is .90 and if an item is dichotomous, then $m = 2$, and the reliability coefficient for that dichotomy is not less than .80:

$$\rho \geqq \frac{2}{1}\left(.90 - \frac{1}{2}\right) = .80$$

If an item is trichotomous, then $m = 3$, and a lower bound to the reliability coefficient for the item is:

$$\rho \geqq \frac{3}{2}\left(.90 - \frac{1}{3}\right) = .85$$

If an item has four categories, then $m = 4$, and a lower bound to the reliability coefficient for such an item is:

$$\rho \geqq \frac{4}{3}\left(.90 - \frac{1}{4}\right) = .867$$

If an item has five categories, then $m = 5$, and a lower bound for ρ is:

$$\rho \geqq \frac{5}{4}\left(.90 - \frac{1}{5}\right) = .875$$

Thus we have an important relationship between the coefficient of reproducibility and the test-retest reliability of each item in the set being analyzed.

It may be noticed that, for fixed reproducibility, the lower bound for an item goes up as the number of categories goes up. But the increment becomes progressively smaller, so that there is not much

difference between the lower bounds of items with four and five categories if they have the same reproducibility.

It is important to remember that only *lower* bounds to the test-retest reliability of the items are being determined this way. *Upper* bounds cannot be established from only one trial. At least two experimentally independent trials are required to set an upper bound.[6] An item may have perfect test-retest reliability for each individual in the population and still have low reproducibility. The property of reproducibility refers to the structure of the universe of items. The property of test-retest reliability refers to the universe of trials of the same individuals on the same item. These are quite distinct concepts and are related only by the inequalities we have just been working with.

The assumptions involved. A basic assumption in the second kind of lower bound is that the item be experimentally independent of the score it is being related to. This will ordinarily not hold if very few items, say only five or six, are analyzed by the scalogram technique. The item itself contributes to the total score and therefore, in general, will have its error of unreliability related to the error of unreliability of the total score. The proper way of using reproducibility as a lower bound for test-retest reliability would be to compute scale scores for individuals on all the items in the sample *except* for the one whose test-retest reliability is being studied for the moment. Then the errors of observation can be experimentally independent. The possible spuriousness of keeping the item in the total score does not seem to be too serious when a dozen or so items are used, if the reproducibility of the whole set is about .90; as we have seen with respect to the sampling of items, it is clear that in such a case scores based on ten items cannot differ very much from those based on eleven items. Since reproducibility is a lower bound anyhow, overestimating the lower bound spuriously by keeping the item in the total score should still yield a lower bound to the reliability coefficient itself, since rarely would we expect the reliability coefficient to be equal to any particular lower bound obtained in practice.

If the items in the set do not have equal reproducibility, then using the overall coefficient of reproducibility as a lower bound does not actually yield a lower bound for each item separately. Instead, it yields a lower bound to the *average* reliability coefficients of the items. If it is desired to establish a lower bound for each item sep-

[6] *Ibid.*

arately, then the actual reproducibility of the item should be used in the formula for lower bound, instead of the overall reproducibility coefficient for the set.

The test-retest reliability of scale scores. Thus far we have investigated the test-retest reliability of responses to separate items. We have seen that this is necessarily very high for approximately scalable items. As a consequence, it should follow that test-retest reliability of scores derived from these items should be very substantial. Again, it should be remembered that we are concerned here with a universe of trials of the same people on the same items; we are not concerned with the problem of sampling of items—that was discussed in a previous part of this chapter. Our problem now is: how much variation will there be in a person's scale score if he were to be given the same items over and over again—without carryover from trial to trial? This is different from the problem previously discussed of how much variation there will be between a person's score obtained on one sample of items and that obtained on another sample of items.

The basic theory and working formulas for this problem of test-retest reliability of scores have been developed recently elsewhere.[7] The problem of quantitative scores is a bit different from that of qualitative items because of a different definition of error. For a qualitative item, we predict the category a person is in either correctly or incorrectly, and amount of error is defined in terms of the total number of incorrect predictions. For quantitative variables, predictions are almost always incorrect; they are rarely exact. But small errors are of no concern; it is the large errors that are important. Hence, error for quantitative variables is measured in terms of the size of errors rather than the number of errors. In the qualitative case, we dealt with modal categories and measured error in terms of frequencies. In the quantitative case, we shall use arithmetic means and measure error in terms of variances.

On each trial in the universe of trials, each person would obtain a total score on the given items. Let e_i^2 be the variance over the trials for the ith person's total score. His score will be said to be completely reliable if his error variance, e_i^2, is zero. The greater e_i^2 is, the more unreliable will the ith person's score be. The problem of test-retest reliability of scores is to obtain information about e_i^2 for each person. To do this exactly would require an actual

[7] Louis Guttman, "A Basis for Analyzing Test-Retest Reliability," *Psychometrika*, Vol. 10, No. 4 (December 1945), pp. 255–282.

series of independent repetitions of the experiment. In practice, it is difficult, if not impossible, to conduct repeated trials without having carryover from trial to trial or else a change in the conditions. It is fortunate that only *a single trial is needed* to provide useful information about the average test-retest reliability of the scores for the population of individuals.

The reliability coefficient for scores. Let ϵ^2 be the arithmetic mean of the e_i^2 for the population. If ϵ^2 is zero, then the population scores will be said to have perfect test-retest reliability; in this case each of the e_i^2 is zero (each individual has zero error variance). The larger ϵ^2 is, the larger are the individual error variances on the average. On the basis of a single trial, it is possible to set an upper bound to ϵ^2.

Instead of working with ϵ^2, however, it is conventional to convert the error variance into a reliability coefficient which varies between zero and unity. For the problem of total scores from a set of items, this coefficient is defined to be:[8]

$$\rho_t^2 = 1 - \frac{\epsilon^2}{\sigma_t^2}$$

where σ_t^2 is the variance of the total scores over the population of individuals and over the universe of trials. It can be proved that ϵ^2 can never be bigger than σ_t^2. If ϵ^2 is zero, then the reliability coefficient, ρ_t^2, is unity; this is the case of perfect test-retest reliability. If ϵ^2 is equal to σ_t^2 (which is its maximum), then the coefficient, ρ_t^2, is zero, which is called the case of perfect unreliability. Intermediate values of ϵ^2 yield intermediate values of ρ_t^2.

The split-half lower bound. On the basis of only a single trial it is possible to set a lower bound to the reliability coefficient, ρ_t^2 (this is equivalent to setting an upper bound to ϵ^2). Six different lower bounds have been developed in the original paper on this subject,[9] some of which are better than others under certain conditions. Perhaps the formula which will give the highest lower bound for scalogram analysis is the split-half lower bound, designated by L_4. (It is the fourth in the series of six bounds.) The computations for the lower bound, L_4, are very simple and will be illustrated for the data on attitude toward officers used in Tables 3 and 5. To compute L_4, it is required that the sample be divided into two halves so

[8] *Ibid.*, p. 262f.
[9] *Ibid.*

that there are two half scores available, as well as the whole score which is the sum of the two half scores. All that is needed, then, is to compute the variances among people of these three sets of scores on a single trial. Let s_a^2 be the variance of the scores in the first half of the test; let s_b^2 be the variance of the scores in the second half of the test; and let s_t^2 be the variance of the total score. Then the formula for the lower bound is:

$$L_4 = 2\left(1 - \frac{s_a^2 + s_b^2}{s_t^2}\right)$$

The variances of the two half scores can be computed from the marginals of Table 5. For the five questions in the first half of the sample, $s_a^2 = 1.99$. For the six questions in the second half, $s_b^2 = 2.35$. The distribution of total scores is given in the marginals of Table 3, whence we ascertain that $s_t^2 = 7.53$. Substituting these three variance values into the formula for L_4, we obtain:

$$L_4 = 2\left(1 - \frac{1.99 + 2.35}{7.53}\right) = .85$$

This means that ρ_t^2, or the reliability coefficient for the total score on the eleven items for the population of soldiers, is *not less* than .85. It may be anything between .85 and unity.

This split-half lower bound, L_4, will tend to be higher according as the correlation between the two half scores increases. Therefore, it is desirable to split the sample into two halves in such a way as to maximize the correlation between the two halves. For a highly reproducible set of items it is clear that two halves can usually be obtained which will have a very substantial intercorrelation, according to our discussion of the problem of sampling of items. The halves should be chosen in such a way as to make their marginal distributions as similar as possible, because, as we have seen previously, two half scores cannot correlate very highly if they have widely different marginals. By pairing off items with similar frequency distributions, and using one from each pair in each half of the sample, it will usually be possible to obtain part scores which have similar distributions and which will have a rather high intercorrelation. This is what was done for the present example on attitude toward officers. As a result, the lower bound to the test-retest reliability coefficient will be high.

The split-half lower bound, L_4, resembles somewhat the traditional Spearman split-half formula, corrected for test length. Indeed, it will be numerically equivalent to Spearman's formula if s_a^2 equals s_b^2. However, it avoids the stringent assumptions behind the Spearman formula, and indeed proves that if Spearman's assumptions are not satisfied by the data, then his formula in general underestimates the test-retest reliability coefficient. The lower bound, L_4, assumes only that the two half sample scores are experimentally independent of each other and that the computations are based on a large population of individuals. It does not assume that the two halves belong to the same scale or anything else of the kind. If the two halves happen to be parts of the same scale, then L_4 will prove to be high. If the two parts do not happen to be scalable together, then L_4 may be low. A low L_4 does *not* imply that the two halves—or their sum, the total score—have low test-retest reliability. The test-retest reliability of the scores may be perfect, yet L_4 can be zero. Unreliability is but one possible reason why variables are uncorrelated. They may be uncorrelated even though they are perfectly reliable.

For the case of the scale scores then, as well as for the qualitative responses to the separate items, we have assurance that if the items are approximately scalable, then they necessarily have very substantial test-retest reliability. Scalogram analysis provides as an automatic byproduct the assurance that responses to individual items and total scores both have relatively little error of measurement if the reproducibility is high.

CHAPTER 9

THE PRINCIPAL COMPONENTS OF SCALE ANALYSIS[1]

The problem of metric. If a set of items conforms to the perfect scale pattern for a population of individuals, then there is a meaningful rank order on those items for the individuals. The rank metric is all that is necessary for one important aspect of external prediction. Any outside variable can be predicted as well from the rank order as it could be from any other scores into which the ranks might be transformed. Therefore, for those problems in which the scale is to be used only as a predictor, there is no need to seek any metric other than the rank order.

On the other hand, if the scale is to be predicted from external variables, and if the prediction is not perfect, then there is some need to consider the problem of metric. The prediction process must minimize errors in some sense. Different metrics will define error differently, so that different metrics may require different prediction processes.

The problem of metric can also arise for internal as well as for external problems. One such internal problem that we have already faced is that of defining a zero point for an attitude. There may be other internal problems arising in the future, and for these it may be desirable to have an intrinsic metric defined for a scale.

The equations of scale analysis to be presented in this chapter show what internal metric, apart from rank order, is involved in the scale pattern. The metric depends on two sets of parameters for the scale: the frequencies of the scale types of persons and the frequencies of the types of scale items. Different frequency distributions give rise to different metrics for scale scores. It is interesting to note that rank order turns out to be a metric arising from a certain pair of frequency distributions.

If we had knowledge of the frequencies of the types of items and

[1] By Louis Guttman.

types of persons, then we would know how to transform the scale ranks into the proper internal metric of the scale, if rank order was itself not the internal metric. Such knowledge seems impossible a priori because of the manner in which items are obtained. Any empirically obtained frequencies are functions of the particular questions constructed by the research worker, and are subject to the vexing difficulties concerning sampling the universe of content (see Chapter 8).

Among the most interesting features of the equations of scale analysis are the new approach they suggest to solving the problem of sampling of items and the new light they throw on how to interpret the content of scales. These are unexpected results that have come out of the quest for an internal metric. The work on the intensity function (see Chapter 7) was one of these results; this work was a direct outcome of the mathematical analysis of the scale pattern. The intensity function is but one of infinitely many functions inherent in a scale; it is the first and only one to date to be identified empirically.[2] These functions are called *principal components*.

Let us see how a scale pattern has inherent in it the intensity and other principal components, and how these relate to the problem of metric.

Description and Illustration of the Scale Equations

For the nonmathematical reader. The notion of principal components is an important one in mathematics and in the physical sciences. It has occurred also in the field of factor analysis for mental testing.[3] In the present chapter we shall show how it

---

[2] Added in proof: The writer has now discovered empirically the next function, or third component of a scale, in his work for the Israel Institute of Public Opinion Research, as reported at the Third Congress of the World Association of Public Opinion Research at Paris in September 1949. This report will be published soon.

The third component is called *closure*, which seems to be a scalable attitude. When related to the content scale of any one attitude of the larger set, it yields a curve with two bends, one on each side of the zero point of content. This can be illustrated from a study of attitudes toward agriculture in Israel. The bending points on definiteness of plans as to farming divided those favorable to agriculture into the more positive and the less positive; the more positive having no stronger alternative than agriculture, the less positive really preferring something else although liking agriculture. Similarly closure separated the more negative from the less negative; those with a definite alternative being the less negative, those with no alternative (that is, unfavorable to agriculture in spite of having no alternative) being the more negative.

[3] See L. L. Thurstone, *The Vectors of Mind* (University of Chicago Press, Chicago, 1935), especially Chapter 4. The most extensive investigation of this problem for factor analysis has been made in Harold Hotelling, "Analysis of a Complex of Statistical Variables into Principal Components," *Journal of Educational Psychology*, Vol. 24, Nos. 6, 7 (September, October 1933), pp. 417–441, 498–520.

emerges naturally, in a special form, for the problem of quantifying qualitative data, with striking results for the scale pattern.

The first half of this chapter is addressed primarily to the nonmathematical reader. In descriptive terms, we shall present the criterion of internal consistency for a metric that leads to the equations that generate the principal components.

For a first reading, it is suggested that the reader not pause very long at any one point until he has worked through all the numerical examples in the first half of this chapter. The actual mathematical derivations and proofs are reserved for the second half.

Scores and weights are not necessary to each other. In quantifying a set of qualitative items, there are two different sets of numerical values that can be determined. One set is the *weights* to be assigned the categories of the items, and the other is the *scores* to be assigned the people. The principle we shall use for quantification, or obtaining a metric, for each of these sets is that of maximizing internal consistency in the sense of least squares. The appropriateness of this criterion for the case of the perfect scale is evidenced by the remarkable results that flow from it.

The most internally consistent scores to assign the people on the basis of their responses to the items are those that satisfy the following condition. All people who fall in one category of an item should have scores as similar as possible among themselves, and as different as possible from the scores of the people in the other categories of the item; this should be true to the best extent possible for all items simultaneously. The total variance of the persons' scores can be expressed as the sum of two parts: the variance of the scores *within* categories, and the variance *between* categories. Maximizing similarities within categories and differences between categories implies maximizing the ratio that the variance between categories has to the total variance. The square root of such a ratio is called technically a *correlation ratio*, so our problem resolves itself into finding the set of scores that has the largest possible correlation ratio with respect to all the items.

Similarly, the most internally consistent values to assign the categories are those that satisfy the following condition. All categories characterizing one person should have numerical values as similar as possible among themselves, and as different as possible from the values of the categories that do not characterize that person; this should be true to the best extent possible for all people simultaneously. The total variance of any numerical values assigned the

categories can also be expressed as the sum of two parts: the variance of the values *within* people, and the variance *between* people. Maximizing similarities within people and differences between people, then, calls for determining the set of numerical values for categories that will have the largest possible correlation ratio with respect to all the people.

It is only to adhere to a conventional terminology, and to help distinguish between values for people and values for categories, that we shall use the term *weight* to refer to a numerical category value.

It is important to observe that if we want only scores for people, *we do not have to pose the problem of weights for categories.* We are *not* specifying in advance that scores are to be obtained by adding up category weights. We specify only the criterion for the scores to satisfy; how to calculate the scores will depend on what the resulting equations show. In fact, the equations for the most internally consistent scores can be set up and solved without ever introducing the notion of any kind of category weights at all. In exactly a similar fashion, the equations for the most internally consistent weights can be set up and solved without ever introducing the notion of any scores at all for people. A metric for people and a metric for categories are two independent problems that can each be solved separately.

The equations of internal consistency. The equations for these two problems were derived over eight years ago.[4] It was also proved that a simple relationship exists between the optimum scores and the optimum weights. The maximum possible correlation ratio for scores is precisely equal to the maximum possible correlation ratio for weights, and a reciprocity exists between the maximizing scores and the maximizing weights. The score of a person is proportional to the arithmetic mean of the weights of the categories by which he is characterized, and the weight of a category is proportional to the arithmetic mean of the scores of the people who are in it.

Hence, if we know the optimum scores, we can calculate the optimum weights by computing arithmetic means of the scores. Conversely, if we knew the optimum weights, we could calculate the optimum scores by computing arithmetic means of these weights.

[4] Louis Guttman, "The Quantification of a Class of Attributes: A Theory and Method of Scale Construction," in P. Horst, et al., *The Prediction of Personal Adjustment* (Social Science Research Council, New York, 1941), pp. 319–348.

Now, each person has as many category weights to average as every other person, because he has one and only one response to each item. Therefore, the arithmetic means of weights are proportional to the sums of the weights, differing by the constant reciprocal of the number of items. As a consequence, the optimum scores are also directly proportional to the *sums* of the optimum weights, or the *scores can be obtained from the weights by essentially an additive process.*

The converse is not true. Weights cannot be obtained from only direct sums of the scores, because the number of people is not constant from category to category. Each category needs its own divisor to obtain the arithmetic mean of scores.

The equations of internal consistency, then, provide justification for assigning weights to items to be *added up* to provide scores for the people. And, what is more important, they show exactly what weights and scores to use if the frequency distributions are known for both the population of individuals and the universe of items.

Scores and weights that "go in a circle." The solution to the equations of internal consistency can be put in the following terms. Find a set of scores for the people such that, if weights are derived from these weights by adding them up for each person, then the new scores will be exactly proportional to the original scores. That is, find a set of scores that "go in a circle": they yield weights that give back the same scores. Such a set of scores is called a principal component[5] of the system.

Or else we could start with the weights. Find a set of weights for the categories such that, if scores are derived by adding them up for each person, and if new weights are derived from these scores by averaging them for each category, then the new weights will be exactly proportional to the original weights. Such weights can also be said to "go in a circle": they yield scores that give back the same weights. Such a set of weights is also called a principal component.

If a set of weights goes in a circle, then the scores obtained from them also go in a circle. Indeed, we might say that it is the same circle involved in both cases. It turns out further that the correlation ratio is the same for scores as for weights if they are in the same circle. To use descriptive language again, we shall say that this common correlation ratio indicates the "size" of the circle.

[5] Principal components are an important concept in mathematics and in physics. It is interesting that they should also arise out of social-psychological considerations. Other names used for them include: principal axes, latent vectors, characteristic functions, characteristic vectors, eigenfunctions, and eigenvectors.

A numerical example. Before going further into properties of principal components, let us consider a numerical example of what it means to "go in a circle." Suppose that a universe of dichotomous items has only five types of items, yielding six types of people.[6] Suppose further that the types of items are all equally frequent, and that the types of persons are all equally frequent. Each type of item represents one fifth of all the items in the universe, and at each scale rank there is one sixth of all the people. The scalogram is as in Table 1.

The two categories of the first type of item are labeled 1a and 1b, and similarly for the other types of items. The types of persons are labeled according to their rank, from five to zero. Since each kind

TABLE 1

SCALE PATTERN OF A UNIVERSE OF FIVE KINDS OF DICHOTOMIES

Type of person	*Type of category*										*Frequency of response*
	1a	*2a*	*3a*	*4a*	*5a*	*1b*	*2b*	*3b*	*4b*	*5b*	
5	x	x	x	x	x						5
4		x	x	x	x	x					5
3			x	x	x	x	x				5
2				x	x	x	x	x			5
1					x	x	x	x	x		5
0						x	x	x	x	x	5
Frequency of response	1	2	3	4	5	5	4	3	2	1	30

of item has the same frequency of occurrence as any other item, and since each kind of person has the same frequency of occurrence as any other kind of person, the relative frequencies of response of people to categories can be ascertained by considering each x in the table to have the same frequency as every other x, that is, each x represents 1/30 of all the responses of the population of people to the universe of the items. (If the types of people and/or the types of items did not have uniform frequencies, the x's in Table 1 would have correspondingly unequal frequencies.)

For this example of uniform frequencies for both people and

[6] Two items are said to be of the same type if they are perfectly correlated with each other. Thus, in a scale, all dichotomies with a 90–10 marginal are of one type, all dichotomies with an 80–20 marginal are of a second type, etc. Frequency of *type* of item, then, is quite different from—and independent of—the frequency distribution of *responses* to the type.

items, we shall verify that the rank order is essentially the intrinsic least squares metric, that is, that rank order scores essentially "go in a circle," and furthermore, have a larger circle or correlation ratio than any other possible set of scores.

First, we shall shift the numbers assigned the ranks so as to make the mean rank zero. As will be explained later, principal component scores must have zero as their mean. For convenience, we shall use whole numbers for the translated ranks, and replace the numbers 5, 4, 3, 2, 1, 0 by the respective numbers 5, 3, 1, −1, −3, −5. These latter numbers we shall prove are a principal component of the scale. By adding 5 to each of these numbers, and then dividing by 2, the original rank numbers are obtained, so the ranks are a perfect linear function of this principal component.

To see that the scores "go in a circle," let us first ascertain the weights they yield for the categories. Category 1a has only one kind of person, whose new score is 5. Hence, the mean of the scores of this kind of person is also 5. Category 2a has two kinds of people, one with score 5 and the other with score 3. Since each type is equally frequent for this example, the mean is 4. Category 3a has the mean score of three kinds of people: $\frac{5+3+1}{3} = 3$, and so on for the rest of the categories. We thus have the following table of weights, or mean scores, for the categories:

TABLE 2

CATEGORY WEIGHTS DERIVED FROM METRIC SCORES

Category	1a	2a	3a	4a	5a	1b	2b	3b	4b	5b
Weight	5	4	3	2	1	−1	−2	−3	−4	−5

A check on the work is that the mean of the weights of each item, taking into account frequency of response, must be zero. Category 1a has a relative response frequency of 1 and category 1b has a relative response frequency of 5. Multiplying the category weights by these frequencies yields the mean weight for items of type 1:

$$(5)(1) + (-1)(5) = 0$$

Similarly, the mean weight for items of type 2 is:

$$(4)(2) + (-2)(4) = 0$$

All the remaining items also have zero mean weights in this fashion.

Now we have to verify that the weights in Table 2 give the metric scores back again. A person in the top rank of Table 1 is in the first five categories; these categories have the respective weights (Table 2) of 5, 4, 3, 2 and 1, so that the mean of the weights for him is:

$$\frac{5 + 4 + 3 + 2 + 1}{5} = 3$$

Each type of item is equally frequent for this example, so that the weights occur equally often for the person and can be averaged in this simple fashion. Similarly, the next highest type of person in Table 1 is in the categories with the respective weights (Table 2) of 4, 3, 2, 1 and −1, so his mean is:

$$\frac{4 + 3 + 2 + 1 - 1}{5} = 1.8$$

The mean of the weights for the third row of Table 1 is:

$$\frac{3 + 2 + 1 - 1 - 2}{5} = .6$$

By taking similar arithmetic means of weights for the remaining people, we obtain scores that can now be compared with the original scores. This is done in Table 3. The new scores just derived as the means of the weights (which weights were in turn derived from the original scores) are called here the "circular" scores. Notice that each circular score is precisely .6 times the original score, or the *circular scores are strictly proportional to the original scores.* Hence

TABLE 3

RELATION BETWEEN ORIGINAL SCORES AND "CIRCULAR" SCORES

Type of person	*Original score*	*"Circular" score*
5	5	3.0
4	3	1.8
3	1	.6
2	−1	− .6
1	−3	−1.8
0	−5	−3.0

these scores do actually go in a circle, and they constitute a principal component of the scale.

It can be proved that the constant of proportionality, which in this case is .6, is always the *square of the correlation ratio* of the component on the items. This we shall call here the "size of circle." It can further be proved that .6 is the largest possible size of circle for our example, and that the scores we are using are the only possible[7] ones that go in this circle. They are the most internally consistent scores in the sense of least squares.

It should be clear that scores that go in a circle are defined only up to a constant of proportionality. Any constant multiples of them will also go in a circle. This is like saying that if scores measured in inches go in a circle, then scores measured in centimeters also go in a circle. For example, if the "circular" scores in Table 3 are now used as original scores, they will also go in a circle. But it will be precisely the same circle as previously, for the new "circular" scores will be exactly .6 times the present "circular" scores.

If new weights are derived from the "circular" scores in Table 3, they will prove to be circular to the old weights of Table 2, being exactly .6 times the previous weights. Hence, the weights of Table 2 also go in a circle of size .6. Any weights proportional to these weights will belong again exactly to this selfsame circle.

There are many principal components for a configuration of responses. It turns out that, for any configuration of responses to items, *whether it is the scale pattern or not,* there exist scores and weights that go in a circle, and hence are the most internally consistent for that configuration. Equally important is the fact that in general there is *more than one* set of scores and of weights which go in a circle, albeit the circles will vary in size. There are many such sets, in fact, usually at least as many as there are distinct types of items. Any such set is also called a principal component of the system. A set of scores that goes in a circle is a principal component of the score system, or of the quantification of persons' ranks; a set of weights that goes in a circle is a principal component of the weight system, or of the quantification of categories values.

Each principal component is associated with a correlation ratio,[8] the square of which is the "size of circle." Hence, of the many principal components, the one with the largest correlation ratio is

[7] All sets of scores proportional to a given set are considered to be essentially the same.

[8] The square of the correlation ratio is called technically a *latent root* of the system. Other names used for this are: characteristic root, characteristic value, and eigenvalue.

the most consistent quantification. The one with the next largest correlation ratio is the most consistent *after* the first component is subtracted out, etc. It is the first component, or the one with the largest correlation ratio, that defines our desired metric; we shall call this the *metric* component.

Examples of further principal components. For our numerical example of Table 1, it can be verified that the following set of scores will also go in a circle: 5, −1, −4, −4, −1, and 5, assigned in this order to the person types. The ranks of these scores are no longer according to the scale rank order. Instead, the scale rank order is "folded over." The persons with the highest and lowest ranks on the scale (types 5 and 0) have the same rank on this second component. Similarly, types 4 and 1 have the next highest rank; and the middle types 3 and 2 have the lowest rank on this second component.

The size of circle now is smaller, for the reader can verify that the constant of proportionality for the "circular" scores obtained from these scores is precisely .2, which is one third of the size of circle (.6) for the metric component. This means that the second component scores by themselves do not discriminate as well between the responses of the people to the items as do the metric component scores. It is easy to see that they could hardly do so in this example, because they implicitly combine the given six types of persons into only three types, according to the folding over of scale ranks. It can be proved, however, that after taking into account the metric component, there is none with a bigger circle than this second one.

Another principal component, which has the third largest possible circle (after the first two are subtracted out), is the following set of scores: 5, −7, −4, 4, 7, −5, assigned to the types of persons in Table 1 in this order. The size of circle here is .1.

The fourth most consistent principal component scores are: 1, −3, 2, 2, −3, 1, with .06 as the size of circle.

The fifth most consistent principal component scores are: 1, −5, 10, −10, 5, −1, with .04 as the size of circle. It can be proved that this also is the *smallest possible circle* for our example, so that this is also the *most in*consistent component. Any other set of numbers, whether a principal component or not, cannot have the square of its correlation ratio less than .04 for five dichotomous and scalable types of items which have uniform frequencies and where the types of persons also have uniform frequencies.

If we add up the five sizes of circles from the five principal components, we obtain unity:

$$.6 + .2 + .1 + .06 + .04 = 1$$

that is, the sum of the squares of the correlation ratios is unity.

The total variation of all the responses in Table 1, expressed as unity, has actually been resolved into precisely the sum of five components. This implies that from the five component scores and weights, we can reproduce every person's answer to every question in Table 1. How to do this will be shown a few sections below.

It can be proved that there is no other set of numbers, except a constant set, that goes in a circle apart from the above five. There are five and only five principal components (and one constant one)

TABLE 4

PRINCIPAL COMPONENT SCORES FOR THE SCALE OF TABLE 1

Type of person	*Principal component score*					
	I	*II*	*III*	*IV*	*V*	*constant*
5	5	5	5	1	1	1
4	3	−1	−7	−3	−5	1
3	1	−4	−4	2	10	1
2	−1	−4	4	2	−10	1
1	−3	−1	7	−3	5	1
0	−5	5	−5	1	−1	1
Size of circle	.6	.2	.1	.06	.04	1
Sum of squares	70	84	180	28	252	6

for a universe of five kinds of dichotomous items. In general, for dichotomous items, the number of nonconstant principal components is equal to the number of types of items.

A constant set of numbers always goes in a circle. Such a set of constants is always proportional to the set of one's: 1, 1, 1, 1, 1, 1. Using these in Table 1 yields weights all equal to unity, which in turn give back "circular" scores of unity, so that the size of circle is itself unity. However, the size of circle here is not the square of a correlation ratio, since the mean of its scores is not zero. All nonconstant principal component scores have zero as their mean. The constant component plays a role analogous to that of the additive constant in a regression equation, and is actually needed later on for our reproducing process to work out exactly. Thus, while

the size of circle for the constant component is the largest possible, it is not to be interpreted as a correlation ratio squared, but we do have to take it into account for completeness later on.

The oscillatory nature of the principal component scores of a scale. Let us bring together the principal component scores of Table 1 into Table 4. The components are labeled with Roman numerals except for the constant one. Column I has the metric component scores; column II has the second component (or what is now identified as intensity) scores, etc.

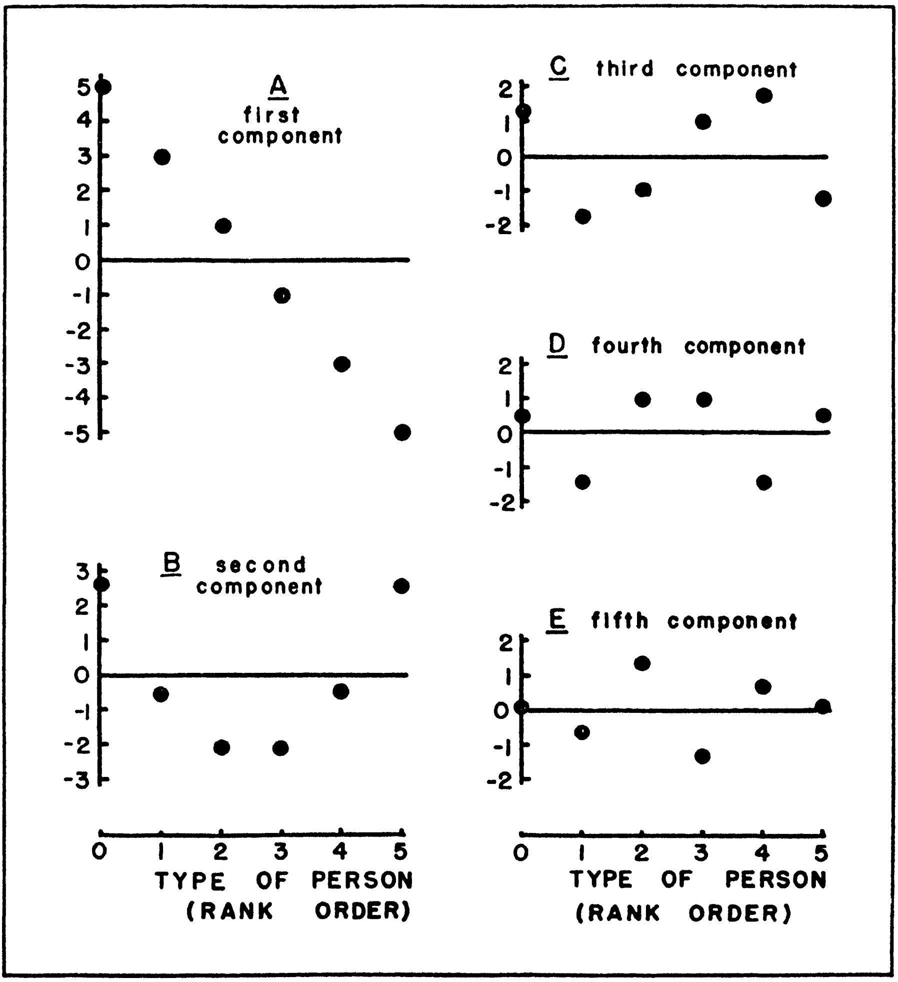


Figure 1. The principal components of the scale of Table 1.

For the perfect scale, such component scores have a definite relation to the rank order of the people. This relationship is shown graphically in Figure 1 for each component separately.[9]

For this example of uniform frequencies for types of persons and for types of items, component I is a perfect straight line function of the rank order (Figure 1a). Hence, rank order is essentially the intrinsic metric for this kind of scalable universe.

The second component is a *U*-shaped function of the rank order for this example (Figure 1b). In fact, the points all lie on a perfect parabola here. The second component is identified psychologically as the intensity component. Since the curve is symmetrical about the middle rank, this indicates that the zero point is between ranks 2 and 3; that is, the zero point separates the component I scores -1 and 1. This is as it should be; the zero point separates positives from negatives on the metric component. If we had an infinite number of types of items instead of just five as in the present example, and if the frequencies remained uniform for both items and people, it can be proved that the infinite number of points for the second component would form a perfectly smooth parabola whose lowest point will be precisely at the zero of the metric scores. The discrete jumps in Figure 1 are due to the fact that we have only a finite number of types in this example.

The curve for the third component (Figure 1c) has one more bend than that for the second component. For the universe of uniform frequencies like our example, it can be proved that the curve is always that of a third degree polynomial. (A parabola represents a polynomial of the second degree.) It could serve to divide the population into three groups, according to its turning points. Perhaps in the future there will be found a psychological interpretation of these bending points, just as the bending point of the parabola

---

[9] The scores in Table 4 are, for simplicity, deliberately expressed in the form of whole numbers, even though this forces the standard deviations to be arbitrarily unequal. It should be remembered that principal components are determined only up to a constant of proportionality: any multiple of a principal component is also a principal component that goes in the same circle and so is not regarded as a new component. This is like changing feet to inches or to centimeters, or pounds to ounces or to grams. The standard deviation changes in direct ratio to the multiplier. It is convenient to use multipliers that will make the standard deviations of the components proportional to their respective correlation ratios. Thus, the metric component—which has the largest size circle—will have the largest standard deviation, while component *V*—which has the smallest size circle—will have the smallest standard deviation. The vertical scales in Figure 1 have been chosen to provide the multipliers needed to make the standard deviations proportional to the respective correlation ratios; the relative sizes of circle are apparent graphically from the average vertical distances of the plotted points from the zero line.

has been identified with the zero point of the attitude or opinion.[10]

Similarly, the fourth and fifth components have curves with three and four bends in them respectively (Figures 1d and 1e). For our example, these represent fourth and fifth degree polynomials. Again, these may be found in the future to provide psychologically meaningful cutting points in the rank order of the people.

It can be proved that there are as many nontrivial principal components for a universe of dichotomies as there are different types of items. Since our example has five types of items, it also has five nontrivial components. Of necessity, there can be at most six distinct metric scores, one for each scale type of person. In order to obtain a continuous metric, there must be an infinite number of types of items and thus also an infinite number of types of people. Then there will also be an infinite number of principal components.

It can be proved that the oscillatory nature of the principal components, as illustrated in Figure 1, also holds if the number of types of items is infinite. The infinite case, with a continuous metric instead of discrete points, is the one usually hypothesized in practice. For this case, if the frequencies remain uniform both for types of items and types of people, then the first, or metric, component will be a *continuous* parabolic function. The third component will be expressible as a *continuous* polynomial of the third degree, etc., etc. Only now there will also be an infinite number of principal components instead of just five as in Figure 1. Each will have one more oscillation in it than the preceding one; for the kind of universe distribution we have hypothesized, each will be expressible precisely as a polynomial of one degree higher than the preceding one.[11]

More generally, the oscillatory nature of the principal components of a scale holds *no matter what the distributions of types of items and types of persons are.* Some types of persons can be more numerous than other types, and/or some types of items can be more numerous than other types of items. The rank scores of people can have a binomial, Poisson, or any other distribution whatsoever as well as the uniform distribution we have used for illustration, and the types of items can also have any arbitrary distribution. Regardless, it will be true that the first component scores will have exactly the same rank order as the rank scores; they will not in general be a

[10] See footnote 2 above added in proof.

[11] The sizes of circles become modified as the number of items changes. In the infinite case, the largest size circle (for the metric component) is .5 for our example; the next largest is .15, etc., etc. The sizes keep decreasing rapidly for the smaller components, but the sum of the infinite series of sizes remains unity, representing the total variation.

straight line function of the rank scores (so that rank scores will not provide the intrinsic metric) but they will be monotonically increasing functions of the rank scores. The first component scores will represent the various stretchings and contractions of distances between ranks needed to provide the intrinsic metric.

The second, or intensity component, will always be *U*- or *J*-shaped. It will not in general be a parabola; it may be part of a sine or cosine curve, or anything else that has one bend in it. It need not be symmetrical about the zero point. If it is symmetrical, that means half the population is positive and half is negative. If it is not symmetrical, then the population is not split 50–50 on the given issue; the majority favors one side.

Similarly, the third component will have two bends in it, even though the equation for it may not be that of a polynomial; the fourth component will have three bends; and so forth.

Components and the sampling of items. The various bending points of the component curves are determined by the respective frequencies of the types of people and types of items. An important feature is that the bending points are essentially invariant with respect to the sampling of items (and of people too); if rank order on a higher component is plotted against scale rank, the bending points will appear at the proper scale ranks. This we have already observed for the second, or intensity, component in Chapter 7. But the same holds true for the remaining components. If a separate measure of the third component could be obtained[12] on a sample of items, like the separate measure we got for intensity in Chapter 7, then plotting the sample ranks on this measure against sample rank on content will yield a curve with two bends in it (instead of the one bend in the intensity curve), and the bends will always occur at the same content percentiles, regardless of the sample of items used. These two bending points of the third component are invariant with respect to sampling of items just like the zero point of the intensity curve. A similar invariance holds for the bending points of the fourth and higher components.

The psychological meaning of the components beyond intensity[13] has not yet been determined, but it seems plausible that each bending point should indicate a psychologically meaningful dividing point for the population, just as the zero point of intensity does. When the psychological meaning is discovered for a component, it

[12] See footnote 2 above added in proof.
[13] See footnote 2 above added in proof.

may be possible then to measure it approximately and independently, as we have intensity.

Because of the invariance of the bending points with respect to sampling of items, they provide a possible solution to the problem of making inferences about the population distribution of scale types and the universe distribution of types of items. We have noted before that these two distributions can be shown by independent experiment to determine the shapes of the component curves (this has been done approximately for the intensity curve). Thus the bending points will provide information about the distributions; such information will not depend on the random sampling of items. The fact that sample items are not random will not prevent valid inferences from them concerning their universe, because of the invariant property of the sample components.

"Scores" and "weights." Thus far, we have concentrated primarily on quantifying the rank order of people. Let us now turn to the quantifying of category values of items. This is a problem distinct from that of people, but it turns out to be closely related. Either of these two problems can be solved separately without reference to the other, but it is convenient for our purposes to present their solutions in a manner which utilizes their interrelationships.

Just as there are values for people that "go in a circle," so there are values for categories that go in a circle. For convenience, we have been calling values for people *scores* here, and values for categories *weights,* in order to distinguish the two quantification problems. It should not be inferred, however, that "weights" are intended a priori only to be added up to provide "scores," any more than "scores" are intended a priori to be added up to provide "weights." It so happens that the equations of internal consistency prove that there is a reciprocal relation between principal components for people and principal components for categories that does involve an additive (more precisely, an *averaging*) process, so we are justified in part in retaining the traditional nomenclature of "scores" and "weights." From the point of view of quantifying people's ranks, this terminology is quite appropriate. From the point of view of quantifying category values, however, the terminology should really be *reversed;* values for people are *weights* which can be used to determine values (scores) for categories. It is convenient not to keep switching terminology, so we shall retain the traditional terms even in the problem of category values.

Example of principal component weights. For each principal com-

ponent of the scores, there is a principal component of the weights with the same correlation ratio and (hence) size of circle. The first principal component weights have already been exhibited in Table 2. If these weights were now used to obtain scores, and the new, "circular," weights were obtained from the scores, the "circular" weights would be found to be precisely .6 times the original weights.

If component weights are known, then component scores can be derived from them by the averaging process. For example, if we did not know the first component scores in Table 4 but did know the first component weights of Table 2, then we could use these weights to obtain the desired scores. Similarly, if the scores are known first, they can be averaged to yield the weights. In our example, we have been given the solution for all the component scores (Table 4). Therefore, we can now easily derive all the component weights from these scores.

The second principal component weights are derived from column II of Table 4, just as the first component weights of Table 2 were derived from the scores in column I of Table 4. The remaining component weights are similarly obtainable from the remaining columns of Table 4. There is again a constant principal component, represented by weights all equal to unity, corresponding to the constant score component. The principal component weights, as obtained from Table 4, are all listed in Table 5.

Like the scores, principal component weights are determined only up to a constant of proportionality. Any column of weights in Table 5 could be multiplied through by any constant whatsoever, and the resulting weights will also go in a circle, but it will remain the same circle as before.

For the perfect scale pattern, each of the remaining component weights has a simple pattern of relationship with the rank of the first, or metric, component weights, analogous to score component relationships. There is a hop in the metric component weights between categories 5a and 1b, where the distance is 2 (the difference between 1 and -1), or twice the distance between any other two neighboring categories. It is convenient, therefore, to consider category 1b as being two ranks below 5a, rather than just one rank. In fact, we could consider the intervening rank to be filled by a category of a sixth type of question, namely, one that has *all* the people in it, for which the weight is zero. Then the metric component will be a perfect straight line function of the category ranks for this example.

The second component is a *U*-shaped function of the category ranks for this example. However, unlike the second component scores, the equation for the *U* is not that of a parabola, but is a more complicated expression. Each of the succeeding component weights exhibits more oscillation than its predecessor in its relationship to the category ranks (the equations for these oscillatory curves differ from those for scores).

Principal components correlate zero with each other. The principal components of a scale present an interesting and important paradox.

TABLE 5

PRINCIPAL COMPONENT WEIGHTS FOR THE SCALE OF TABLE 1

Type of category	*Principal component weight*						*Frequency of response*
	I	*II*	*III*	*IV*	*V*	*Constant*	
1a	5	5	5	1	1	1	1
2a	4	2	−1	−1	−2	1	2
3a	3	0	−2	0	2	1	3
4a	2	−1	−0.5	0.5	−1	1	4
5a	1	−1	1	−0.2	0.2	1	5
1b	−1	−1	−1	−0.2	−0.2	1	5
2b	−2	−1	0.5	0.5	1	1	4
3b	−3	0	2	0	−2	1	3
4b	−4	2	1	−1	2	1	2
5b	−5	5	−5	1	−1	1	1
Size of circle	.6	.2	.1	.06	.04	1	
					Total frequency		30

Although the principal component scores are all perfect functions of the scale ranks, nevertheless *they have perfect zero linear correlations with each other.* Let us verify this for the scores in Table 4.

To show that a correlation coefficient is zero, it is sufficient merely to show that its numerator is zero. Since the principal component scores have zero means, the numerator of the product-moment correlation coefficient can be obtained by taking merely the products of the scores as they stand. For example, the numerator for the linear correlation between component scores I and II is computed as follows. The first type of person has a score of 5 on both of these components, so the product of his scores is 25. A person of type 4 has a score of 3 on the first component and −1 on the second component, so his product is −3. The products for the remaining types are similarly computed. The product sum, or numerator for the linear

correlation coefficient, between components I and II is found to be precisely zero:

$$(5)(5) + (3)(-1) + (1)(-4) + (-1)(-4) + (-3)(-1) + (-5)(5) = 0$$

This means that the correlation coefficient itself is zero. Similarly, correlations between any two columns of scores in Table 4 can be verified to be zero. Thus we have the paradox of uncorrelated but dependent sets of scores.

Similarly, it can be verified that the principal component weights of Table 5 have zero linear correlations. Here we have to take into account the fact that the categories have different frequencies of response. To calculate the numerator of the correlation coefficient between weights of the first and second components, we proceed as follows. Category 1a has a weight of 5 on both components and a frequency of 1, so its product is $(5)(5)(1) = 25$. Category 2a has a weight of 4 in the first component and a weight of 2 in the second component, with a frequency of 2, so its product is $(4)(2)(2) = 16$. The products for the remaining categories are computed in a similar fashion, and the sum of the products is precisely zero:

$$(5)(5)(1) + (4)(2)(2) + (3)(0)(3) + (2)(-1)(4) + (1)(-1)(5) + (-1)(-1)(5) + (-2)(-1)(4) + (-3)(0)(3) + (-4)(2)(2) + (-5)(5)(1) = 0$$

In a similar fashion the correlation between any two columns in Table 5 can be seen to be zero.

It is this property of zero linear correlations but perfect curvilinear correlations that will throw off any attempt at factor analysis in the Spearman-Thurstone sense. The Spearman-Thurstone approach was devised for a different kind of data; it is not equipped to handle configurations of the kind involved in scalogram analysis.

It is of great psychological importance that we find many linear dimensions in a scale, even though there is but a single factor. The response of each individual to each item can be viewed in either of two ways. The first way is the one thus far used in the preceding chapters: the responses are a simple function of the rank order of the people. The second way of looking at the matter is the one that has now arisen out of our quest for an internal metric: the responses to an item are expressible as functions of metric scores, intensity scores, and scores on principal components of higher order.

The responses have been resolved into many linear components, even though they are functions of but a single factor. Each item involves the same single factor as any other item, but differs in how it combines the components. Each type of item represents a different variation on the same theme. Another way of expressing this is that each type of item represents a different combination of the tones and overtones of the same basic note.

Different scalable attitudes will have different metrics, reflecting the differences in the shapes of the several component curves. Different intensity structures of attitudes imply different metrics, as will different structures of lower order components.

Thus, we have further evidence of the appropriateness of scalogram analysis to the study of attitudes and opinions in that it shows the deep relationship that exists between unidimensionality and the concomitant multiplicity of psychological components.

Reproducing responses from principal components. Thus far, we have shown how the principal components are involved in and derivable from the item responses. Now we shall show how the item responses can be reproduced from the principal components.

Just as responses of individuals to items can be reproduced from the fact that they are simple functions of the scale ranks, so they can be reproduced from knowledge of their principal components. The procedure for doing the reproducing is different for these two ways of viewing the matter. Using the notion of simple function, we establish cutting points on the rank order for each item and thereby determine whether or not a person falls in a given category. To reproduce the responses from the principal components, a simple numerical scheme holds. We can reproduce the responses of Table 1 by means of product sums of principal component weights and scores.

The numerical procedure can be illustrated as follows. Suppose we want to know whether a person of type 5 falls in category 2a. We know the five component scores for such a person from the first row of Table 4: 5, 5, 5, 1, and 1. He also has a constant score of 1. From Table 5, we know the component weights for category 2a: 4, 2, −1, −1, −2, and the constant weight of 1. Now we have to remember that the scores and weights are determined only up to a constant of proportionality; they must be reduced to a standardized form for the purpose of reproduction. The standardizing is done with the aid of the sum of squares of the columns in Table 4, which are indicated in the last row of Table 4. A person of type 5 has a

score of 5 on component I, while category 2a has a weight of 4 on this component, and the sum of squares is 70. The standardized product of score and weight for the first component then is: (5)(4)/70. The standardized product of score and weight for the second component is: (5)(2)/84, and so on for the remaining three components, and also for the constant component. These six standardized products are added up and then multiplied by the frequency of response of the category, which in this case is 2 (from the last column of Table 5 or the last row of Table 1). The final result turns out to be exactly unity:

$$2\left[\frac{(5)(4)}{70} + \frac{(5)(2)}{84} + \frac{(5)(-1)}{180} + \frac{(1)(-1)}{28} + \frac{(1)(-2)}{252} + \frac{(1)(1)}{6}\right] = 1$$

The result will always be unity if the person falls in the category, and it will be zero if the person does not fall in the category. For each x in Table 1 we would obtain unity as the outcome of the standardized sum of the products, and where there is no x we would obtain zero.

As an example of obtaining a zero, let us inquire if a person of type 5 is in category 3b. This category has a frequency of response of 3, and if we multiply this by the sum of standardized products we obtain zero:

$$3\left[\frac{(5)(-3)}{70} + \frac{(5)(0)}{84} + \frac{(5)(2)}{180} + \frac{(1)(0)}{28} + \frac{(1)(-2)}{252} + \frac{(1)(1)}{6}\right] = 0$$

Hence, a person of type 5 does not fall in category 3b; there actually is no x in Table 1 for that position.

We could test every category for a person of type 5 and we would find the standardized product sums to yield unity or zero in the appropriate places, and similarly for persons of type 4 and the other types. The pattern of 1's and 0's obtained from the principal components is shown in Table 6. This is identical in form with Table 1. The principal components reproduce the original responses exactly, just as do cutting points on the scale rank order.

The distinction between scalability and nonscalability. Every pattern of responses, whether scalable or not, can be resolved into principal components, and the responses can be reproduced exactly from all the components. The distinction between scalable and nonscalable configurations lies in the *structure* of the principal com-

ponents. Just as it is true for nonscalable data that they cannot be reproduced perfectly from cutting points on any rank order of people, so it is true that the principal components are *not* perfect functions of any rank order. For a scalable configuration, we have observed the neat relationships among the hierarchy of components. All the components are perfect (curvilinear) functions of the rank order (and of the first or metric component); and they have a definite law of formation as to their oscillations. This is not at all true for nonscalable data. There is no perfect curvilinear relationship between the components. Indeed, they are always

TABLE 6

RESPONSES OF TABLE 1 REPRODUCED FROM THE PRINCIPAL COMPONENTS

Type of person	*Type of category*									
	1a	*2a*	*3a*	*4a*	*5a*	*1b*	*2b*	*3b*	*4b*	*5b*
5	1	1	1	1	1	0	0	0	0	0
4	0	1	1	1	1	1	0	0	0	0
3	0	0	1	1	1	1	1	0	0	0
2	0	0	0	1	1	1	1	1	0	0
1	0	0	0	0	1	1	1	1	1	0
0	0	0	0	0	0	1	1	1	1	1

uncorrelated just as in the scalable case, but furthermore they have no simple pattern of dependence and can actually be completely independent. In general, there is no such thing as a U-shaped second component which might be identified as intensity, nor a third component which has two bends, etc., etc. From the first component one cannot know the other components in the case of nonscalable data.

In the scalable case there seems to be hope that the principal components can be identified experimentally, as we have done with the intensity function. This might serve as a solution to the problem of sampling of items. In the nonscalable case there seems to be no solution as yet in sight for the problem of sampling items. There is as yet no framework of inference available concerning the relationship between a sample of items and the universe of items for the nonscalable case, nor any clues as to how to identify or interpret the components. This again is but a reflection of the fact that there is no single meaningful rank order for the general nonscalable case.

The Mathematical Analysis

The parallelogram for dichotomies. Let us consider a scalable universe of m types of dichotomous items, so that there are $m + 1$ types of persons. The types of persons will be numbered in their scale order from zero to m. There are $2m$ types of categories (two for each type of item); these will be numbered according to their scale order from one to $2m$. Then categories j and $m + j$ $(j \leqq m)$ belong to the same dichotomy, or we shall say they comprise the jth type of dichotomy.

TABLE 7

MATRIX OF A SCALE OF m TYPES OF DICHOTOMIES

Type of person	*Type of category*						*Frequency of response*	*Frequency of person type*
	1	*2*	*3* $\cdots$ *m*	$m+1$	$m+2$	$m+3 \cdots 2m$		
0	1	1	$1 \cdots 1$	0	0	$0 \cdots 0$	n	f_0
1	0	1	$1 \cdots 1$	1	0	$0 \cdots 0$	n	f_1
2	0	0	$1 \cdots 1$	1	1	$0 \cdots 0$	n	f_2
...	..	..	....	..	......	......		...
$m - 1$	0	0	$0 \cdots 1$	1	1	$1 \cdots 0$	n	f_{m-1}
m	0	0	$0 \cdots 0$	1	1	$1 \cdots 1$	n	f_m
Frequency of response	F_1	F_2	$F_3 \cdots F_m$	F_{m+1}	F_{m+2}	$F_{m+3} \cdots F_{2m}$		
Frequency of category types	g_1	g_2	$g_3 \cdots g_m$	g_{m+1}	g_{m+2}	$g_{m+3} \cdots g_{2m}$		

The response of each type of person to each type of item will be represented compactly as follows. Let

$$(1) \qquad e_{ij} = \begin{cases} 1 & \text{if persons of type } i \\ & \text{are in category type } j, \\ 0 & \text{otherwise,} \end{cases} \qquad \begin{pmatrix} i = 0, 1, \cdots, m \\ j = 1, 2, \cdots, 2m \end{pmatrix}$$

The scale matrix defined by (1) is exhibited by the ones and zeros in the body of Table 7.

The parallelogram pattern of ones and zeros can be stated in more analytical terms:

$$(2)\qquad e_{ij} = \begin{cases} 1 & \text{if } 1 \leqq j - i \leqq m, \\ \\ 0 & \text{otherwise,} \end{cases} \qquad \begin{pmatrix} i = 0, 1, \cdots, m \\ j = 1, 2, \cdots, 2m \end{pmatrix}$$

It follows that, if $e_{ij} = 1$, then $e_{i,m+j} = 0$, and conversely, so that:

$$(3)\qquad e_{ij} + e_{i,m+j} = 1, \qquad \begin{pmatrix} i = 0, 1, \cdots, m \\ j = 1, 2, \cdots, m \end{pmatrix}$$

This states that if a person is in category j, he cannot be simultaneously in category $m + j$ (which is the opposite category of the same item); conversely, if he is in category $m + j$, he cannot be in category j; he must be in one and only one of these two categories.

Frequencies of types and of responses. The parallelogram pattern shows how types of persons behave with respect to types of categories. We have also to consider the frequency distributions of these two kinds of types. Let f_i be the number of persons of the ith type ($i = 0, 1, \cdots, m$), and let N be the total number of people in the population. Then

$$(4)\qquad N = \sum_{i=0} f_i$$

Let g_j be the number of categories of the jth type. The f_i and g_j are indicated in the outer margins of Table 7. Since categories belonging to the same type of item occur equally often in the universe, we have

$$(5)\qquad g_j = g_{m+j} \qquad (j = 1, 2, \cdots, m)$$

Let n be the total number of items in the universe. Then

$$(6)\qquad n = \sum_{j=1}^{m} g_j$$

Next, we consider the frequencies of responses. Since each person is in one and only one category of each item, and since there are n items, each individual has n responses. Analytically, this is expressed as follows. If $e_{ij} = 1$, then a person of type i has g_j responses of the category j type. Therefore, his total number of responses is obtained by adding up those g_j for which $e_{ij} = 1$ for him. If we call this sum G_i, then

$$(7) \qquad G_i = \sum_{j=1}^{2m} g_j e_{ij} = \sum_{j=1}^{m} g_j e_{ij} + \sum_{j=1}^{m} g_{m+j} e_{i,m+j} \qquad (i = 0, 1, \cdots, m)$$

From (5), g_{m+j} can be replaced by g_j in the second term on the right. Then using (3) and (6) reduces the whole right member of (7) to n, so that we have the identity:

$$(8) \qquad n = \sum_{j=1} g_j e_{ij} \qquad (i = 0, 1, \cdots, m)$$

The frequency of response for categories, however, is not constant like that for people. Each type of category can have a different number of people in it from any other type of category. Let F_j be the number of people in—or the frequency of response to—a category of the jth type. Then

$$(9) \qquad F_j = \sum_{i=0}^{m} f_i e_{ij} \qquad (j = 1, 2, \cdots, 2m)$$

The two categories of an item must contain all the N people, because each person is in one of the categories of each dichotomy. This is seen analytically by the fact that, from (9),

$$(10) \qquad F_{m+j} = \sum_{i=0}^{m} f_i e_{i,m+j} \qquad (j = 1, 2, \cdots, m)$$

so that adding right members of (9) and (10) and then using (3) and (4) yields the identity:

$$(11) \qquad F_j + F_{m+j} = N, \qquad (j = 1, 2, \cdots, m)$$

The frequencies of response are shown in the inner margins of Table 7.

The correlation ratios. The two problems of quantification, that of finding the most internally consistent numerical values for people and the similar problem for categories, can be posed and solved separately. Numerical values for categories will be called "weights" to distinguish them from the numerical values for persons which are called "scores." No implication of a "weighing" or a "weighting" process is intended here a priori.

The basic equations for these two problems have been developed previously,[14] for the case more general than the scale pattern. For

[14] Guttman, *loc. cit.*

the special case of a scale, we obtain very striking and important results. The basic equations will now be repeated here, in the notation of the present chapter, so that the discussion here will be relatively self-contained.

For the problem of scores, we have to consider the variation within *columns* of Table 7. For a given set of x_i, the mean of scores for the jth column or type of category is a_j, where

$$a_j = \frac{1}{F_j} \sum_{h=0}^{m} x_h f_h e_{hj} \qquad (j = 1, 2, \cdots, 2m)$$

The sum of squares of deviations from the mean for the jth column is s_j, where

$$s_j = \sum_{i=0}^{m} (x_i - a_j)^2 f_i e_{ij} \qquad (j = 1, 2, \cdots, 2m)$$

The total sum of squares over all columns, or the sum of the variation *within* all categories, is S_x, where

$$S_x = \sum_{j=1}^{2m} s_j g_j$$

We wish to find the x_i which will minimize S_x as compared to the total variation about the general mean.

The general mean of the x_i can be denoted by A, where

$$A = \frac{1}{N} \sum_{h=0}^{m} x_h f_h$$

The total sum of squares of the deviations of the x_i from A, summing first over i and then over j, is seen to be T_x, where

$$T_x = n \sum_{i=0}^{m} (x_i - A)^2 f_i$$

We wish to minimize the ratio S_x/T_x. Now the complement of this ratio is the square of the *correlation ratio* for scores for Table 7, denoted by η_x. The problem is thus equivalent to maximizing η_x, where

$$\eta_x^2 = 1 - S_x/T_x$$

Therefore, in expanded notation, for the problem of scores, the best for reproducing e_{ij} is that set of values x_i that will maximize the correlation ratio η_x defined by

$$(12) \qquad \eta_x^2 = 1 - \frac{\sum_{i=0}^{m} \sum_{j=1}^{2m} (x_i - \frac{1}{F_j} \sum_{h=0}^{m} x_h f_h e_{hj})^2 f_i g_j e_{ij}}{n \sum_{i=0}^{m} (x_i - \frac{1}{N} \sum_{h=0}^{m} x_h f_h)^2 f_i}$$

The maximizing scores satisfy the stationary equations obtained by differentiating (12) with respect to the x_i and setting the derivatives equal to zero:

$$(13) \qquad \sum_{h=0}^{m} x_h f_h A_{hi} = \eta_x^2 x_i \qquad (i = 0, 1, \cdots, m)$$

where $\| A_{hi} \|$ is the Gramian matrix defined by

$$(14) \qquad A_{hi} = \frac{1}{n} \sum_{j=1}^{2m} \frac{g_j e_{hj} e_{ij}}{F_j} \qquad (h, i = 0, 1, \cdots, m)$$

Similarly, the problem of weights is concerned with minimizing variation within *rows* of Table 7. The correlation ratio for rows is determined as follows. For a given set of weights, y_j, the mean for the ith row of Table 7 is b_i, where

$$b_i = \frac{1}{n} \sum_{k=1}^{2m} y_k g_k e_{ik} \qquad (i = 0, 1, \cdots, m)$$

The sum of squares of deviations from the mean for the ith row is r_i, where

$$r_i = \sum_{j=1}^{2m} (y_j - b_i)^2 g_j e_{ij} \qquad (i = 0, 1, \cdots, m)$$

The total sum of squares within rows, that is, within people, is then R_y, where

$$R_y = \sum_{i=0}^{m} r_i f_i$$

The general mean for the weights is B, where

$$B = \frac{1}{Nn} \sum_{k=1}^{2m} y_k F_k$$

and the total variation around B for all rows is U_y, where

$$U_y = \sum_{j=1}^{2m} (y_j - B)^2 F_j$$

The ratio to be minimized is R_y/U_y, or the correlation ratio to be maximized is η_y, where

$$\eta_y^2 = 1 - R_y/U_y$$

Therefore, for the problem of weights, in expanded notation, the best ones are the set of values y_j that will maximize the correlation ratio η_y defined by

$$\eta_y^2 = 1 - \frac{\sum_{i=0}^{m} \sum_{j=1}^{2m} (y_j - \frac{1}{n} \sum_{k=1}^{2m} y_k g_k e_{ik})^2 f_i g_j e_{ij}}{\sum_{j=1}^{2m} (y_j - \frac{1}{Nn} \sum_{k=1}^{2m} y_k F_k)^2 F_j} \tag{15}$$

for which the stationary equations are

$$\sum_{j=1}^{2m} y_j g_j B_{jk} = \eta_y^2 F_k y_k \qquad (k = 1, 2, \cdots, 2m) \tag{16}$$

where $\| B_{jk} \|$ is the Gramian matrix defined by

$$B_{jk} = \frac{1}{n} \sum_{i=0}^{m} f_i e_{ij} e_{ik} \qquad (j, k = 1, 2, \cdots, 2m) \tag{17}$$

The equations of internal consistency. As has been shown elsewhere,[15] the optimum scores satisfying (13) and the optimum weights satisfying (16) are related. They together also solve a third quantification problem that can be posed: find scores and weights that are most consistent with each other. This requires finding the values x_i and y_j simultaneously that will maximize the correlation coefficient ρ defined as

(18) $\rho =$

$$\frac{\sum_{i=0}^{m} \sum_{j=1}^{2m} (x_i - \frac{1}{N} \sum_{h=0}^{m} x_h f_h)(y_j - \frac{1}{Nn} \sum_{k=1}^{2m} y_k F_k) f_i g_j e_{ij}}{\sqrt{\left[n \sum_{i=0}^{m} (x_i - \frac{1}{N} \sum_{h=0}^{m} x_h f_h)^2 f_i\right] \left[\sum_{j=1}^{2m} (y_j - \frac{1}{Nn} \sum_{k=1}^{2m} y_k F_k)^2 F_j\right]}}$$

[15] *Ibid.*

The stationary equations are:

(19) $$\sum_{j=1}^{2m} y_j g_j e_{ij} = \lambda n x_i \qquad (i = 0, 1, \cdots, m)$$

(20) $$\sum_{i=0}^{m} x_i f_i e_{ij} = \mu y_j F_j \qquad (j = 1, 2, \cdots, 2m)$$

where

(21) $$\lambda\mu = \rho^2$$

Equations (19) and (20) have been called the *equations of internal consistency*. They are equivalent to both (13) and (16); using (20) in (19) to eliminate the y_j yields (13), and using (19) in (20) to eliminate the x_i yields (16). The equations of internal consistency state that, for stationary values, the score of each person is proportional to the mean of the weights of the categories he falls in, and the weight of each category is proportional to the mean of the scores of the persons who fall in that category.

Latent roots and vectors. Because of the equivalence of the equations, a solution of (19) and (20) with correlation coefficient ρ must yield correlation ratios for (13) and (16) such that

(22) $$\rho^2 = \eta_x^2 = \eta_y^2$$

and this must be true for each solution.

The equations have more than one solution. Equations (13) and (16) can be regarded as the respective characteristic equations of the matrices $\| f_h A_{hi} \|$ and $\| g_j B_{jk}/F_k \|$. The x_i are the elements of a latent vector of the first matrix, corresponding to the latent root η_x^2; and the y_j are the elements of a latent vector of the second matrix, corresponding to the latent root η_y^2. The ranks of the two matrices are the same; they have the same number of nonzero latent roots—all of which are *positive*—and these roots satisfy (22).

There are always the constant latent vectors $x_i \equiv y_j \equiv 1$, for which the latent roots are unity; but these latent roots are not actually squares of correlation ratios. All nonconstant latent vectors have roots less than unity (these roots really being squares of correlation ratios), and these vectors satisfy the respective conditions

(23) $$\sum_{i=0}^{m} f_i x_i = 0$$

$$F_j y_j + F_{m+j} y_{m+j} = 0, \qquad (j = 1, 2, \cdots, m) \tag{24}$$

It is the largest root of the nonconstant latent vectors that is the maximum ρ^2 (or η_x^2 or η_y^2), and the respective corresponding latent vectors—which will be proved to be unique in general for our scale pattern—are the optimum scores and weights, or what we call here the *metric* components.

For further discussion of the solutions of the matric equations for an arbitrary pattern of responses, scalable or not, the reader is referred to the original paper.[16] We continue now with the special form the solution takes for the dichotomous scale pattern. In this special case, the matric equations transform into an important—and apparently new—general type of *difference* equations. In this case, the remaining latent vectors are of considerable interest, as well as the metric components.

The first differences. The matric equations for the dichotomous scale transform into a second-order linear difference equation that does not seem to have appeared before in its general form, and that is of some general mathematical importance apart from scale analysis. Both equations (13) for scores and (16) for weights yield this form of equation, differing in the coefficients and boundary conditions. It is perhaps easier to work with (19) and (20), which are equivalent to (13) and (16), and that is where we shall begin the differencing.

Taking the first difference of both members of (19) with respect to i yields

$$\lambda n \Delta_i x_i = \sum_{j=1}^{2m} y_j g_j \Delta_i e_{ij} \qquad (i = 0, 1, \cdots, m-1) \tag{25}$$

From (2),

$$\Delta_i e_{ij} = \begin{cases} 1 & j = i + m + 1, \\ -1 & j = i + 1, \\ 0 & \text{otherwise}, \end{cases} \qquad \begin{pmatrix} i = 0, 1, \cdots, m-1 \\ j = 1, 2, \cdots, 2m \end{pmatrix} \tag{26}$$

Using (26), (25) becomes

$$\lambda n \Delta_i x_i = y_{i+m+1} g_{i+m+1} - y_{i+1} g_{i+1} \qquad (i = 0, 1, \cdots, m-1) \tag{27}$$

or, remembering (5),

[16] *Ibid.*

(28) $\lambda n\Delta_i x_i = g_{i+1}(y_{i+m+1} - y_{i+1}) \qquad (i = 0, 1, \cdots, m-1)$

But, from (24), replacing j by $i + 1$,

(29) $y_{i+m+1} = -F_{i+1}y_{i+1}/F_{i+m+1} \qquad (i = 0, 1, \cdots, m-1)$

Using (29) in (28), and remembering (11), we obtain

(30) $$\lambda n\Delta_i x_i = \frac{-Ng_{i+1}}{F_{i+m+1}} y_{i+1} \qquad (i = 0, 1, \cdots, m-1)$$

Next we operate on the y_j. Taking the first difference of both members of (20) with respect to j yields

(31) $$\mu\Delta_j F_j y_j = \sum_{i=0}^{m} x_i f_i \Delta_j e_{ij} \qquad (j = 1, 2, \cdots, 2m-1)$$

From (2),

(32) $$\Delta_j e_{ij} = \begin{cases} 1 & i = j \\ -1 & i = j - m, \\ 0 & \text{otherwise,} \end{cases} \qquad \begin{pmatrix} i = 0, 1, \cdots, m \\ j = 1, 2, \cdots, 2m-1 \end{pmatrix}$$

Using this result in (31), we find

(33) $$\mu\Delta_j F_j y_j = \begin{cases} f_j x_j & (1 \leqq j \leqq m-1) \\ -f_{j-m}x_{j-m} & (m+1 \leqq j \leqq 2m-1) \\ f_m x_m - f_0 x_0 & (j = m) \end{cases}$$

The difference equation for scores. For the second differences it is convenient to introduce the following notation. Let

(34) $$c_i = \frac{F_{i+1}F_{i+m+1}}{g_{i+1}} \qquad (i = 0, 1, \cdots, m-1)$$

and let

(35) $$\phi = N/(n\lambda\mu)$$

Then (30) can be written as

(36) $$c_i\Delta_i x_i = -\phi\mu F_{i+1}y_{i+1} \qquad (i = 0, 1, \cdots, m-1)$$

Taking the first difference of both members of (36) with respect to i yields

(37) $$\Delta_\iota(c_\iota\Delta_\iota x_i) = -\phi\mu\Delta_\iota F_{i+1}y_{\iota+1} \qquad (i = 0, 1, \cdots, m-2)$$

whence, using (33), we have the desired second-order equation in the x_ι alone:

(38) $$\Delta_\iota(c_\iota\Delta_\iota x_\iota) + \phi f_{\iota+1}x_{\iota+1} = 0,^{\cdot} \qquad (i = 0, 1, \cdots, m-2)$$

The boundary conditions for (38) follow from the scale pattern. According to (2),

(39) $$e_{\iota 1} = \begin{cases} 1, & (i = 0) \\ 0, & (1 \leqq i \leqq m) \end{cases}$$

that is, the first category column of Table 7 has unity for a person of the lowest type only. Therefore, from (39) and (20),

(40) $$\mu y_1 F_1 = f_0 x_0$$

Using (40) in (36), we obtain a lower boundary condition:

(41) $$c_0\Delta_0 x_0 + \phi f_0 x_0 = 0$$

Similarly, to obtain an upper boundary condition, we notice that from (2),

(42) $$e_{\iota,2m} = \begin{cases} 1, & (i = m) \\ 0, & (0 \leqq i \leqq m-1) \end{cases}$$

that is, the last category column of Table 7 has unity for a person of the highest type only. Using (42) in (20) yields

(43) $$\mu y_{2m} F_{2m} = f_m x_m$$

Using (24) and (43) in (36), we obtain an upper boundary condition:

(44) $$c_{m-1}\Delta_{m-1}x_{m-1} - \phi f_m x_m = 0$$

The difference equation for weights. An equation of precisely the same form as (38) holds for weights, with boundary conditions of a rather different form from (41) and (44). These are obtainable from (33). We need concern ourselves only with the first m weights,

for the remaining weights are then easily obtainable by virtue of (24).

According to (33) and the notation (35),

$$\text{(45)}\qquad \frac{N}{\phi f_j}\Delta_j F_j y_j = \lambda n x_j \qquad (j = 1, 2, \cdots, m-1)$$

Taking the first difference of each member of (45) with respect to j, and then using (30) and (34), yields the desired difference equation in the y_j alone:

$$\text{(46)}\quad \Delta_j\left(\frac{1}{f_j}\Delta_j F_j y_j\right) + \frac{\phi}{c_j}F_{j+1}y_{j+1} = 0, \qquad (j = 1, 2, \cdots, m-2)$$

This is of the same form as (38), with the $F_j y_j$, $1/f_j$, and $1/c_j$ playing the roles of the x_i, c_i, and f_{i+1} respectively.

The boundary conditions for (46), however, have a different form from those of (38). The lower boundary condition is obtained by solving (41) for x_1 in terms of x_0, and then in terms of y_1 by means of (40):

$$\text{(47)}\qquad x_1 = \frac{\mu F_1 y_1(c_0 - \phi f_0)}{c_0 f_0}$$

Then using (47) in (33), we arrive at the boundary condition

$$\text{(48)}\qquad \frac{1}{f_1}\Delta_1 F_1 y_1 + \left(\frac{\phi}{c_0} - \frac{1}{f_0}\right)F_1 y_1 = 0$$

For the upper boundary condition, we solve (44) for x_{m-1} in terms of x_m, and then we use (43) and (24) to obtain

$$\text{(49)}\qquad x_{m-1} = \frac{\mu F_m y_m(\phi f_m - c_{m-1})}{c_{m-1} f_m}$$

Using (49) in (33) yields the boundary condition

$$\text{(50)}\qquad \frac{1}{f_{m-1}}\Delta_{m-1}F_{m-1}y_{m-1} - \left(\frac{\phi}{c_{m-1}} - \frac{1}{f_m}\right)F_m y_m = 0$$

Conditions (48) and (50) differ in form from (41) and (44) in that the coefficients of the last terms have an additive constant.

Matric inversion by differencing. The difference equation (38) and its boundary conditions (41) and (44) are equivalent to the

matric equation (13). Any solution of (13) necessarily satisfies (38), provided only that it also satisfies (24); this follows from the derivation of (28). Only the constant solution does not satisfy (24); it is easily verified, however, that it in fact does satisfy (38), only that now $1/\lambda\mu$ equals zero instead of unity for the solution.

The converse, that every solution of (38) also satisfies (13), can be seen by regarding the difference equations as a matric equation. Equations (41), (38), and (44) can be written in the respective expanded forms:

$$x_0c_0 - x_1c_0 = \phi f_0 x_0 \tag{51}$$

$$-x_ic_i + x_{i+1}(c_i + c_{i+1}) - x_{i+2}c_{i+1} = \phi f_{i+1}x_{i+1} \qquad (i = 0, 1, \cdots, m-2) \tag{52}$$

$$-x_{m-1}c_{m-1} + x_mc_{m-1} = \phi f_m x_m \tag{53}$$

Let C_{hi} $(h, i = 0, 1, \cdots, m)$ be the elements of the symmetric matrix C defined as

$$(54)\ C = \begin{Vmatrix} c_0 & -c_0 & & & & & \\ -c_0 & c_0 + c_1 & -c_1 & & & & \\ & -c_1 & c_1 + c_2 & -c_2 & & & \\ & & -c_2 & c_2 + c_3 & -c_3 & & \\ & & & \cdots\cdots & & & \\ & & & & -c_{m-2} & c_{m-2} + c_{m-1} & -c_{m-1} \\ & & & & & -c_{m-1} & c_{m-1} \end{Vmatrix}$$

Then equations (51), (52), and (53) can be written as the single matric equation

$$\sum_{h=0}^{m} x_hC_{hi} = \phi f_i x_i \qquad (i = 0, 1, \cdots, m) \tag{55}$$

In order to help compare (55) with (13), it is convenient to restate both these equations as follows. Let

$$\bar{x}_i = x_i\sqrt{f_i} \qquad (i = 0, 1, \cdots, m) \tag{56}$$

$$\bar{A}_{hi} = A_{hi}\sqrt{f_hf_i} \qquad (h, i = 0, 1, \cdots, m) \tag{57}$$

$$\bar{C}_{hi} = \frac{nC_{hi}}{N\sqrt{f_hf_i}} \qquad (h, i = 0, 1, \cdots, m) \tag{58}$$

and let

$$\bar{x} = \|\bar{x}_i\|, \qquad \bar{A} = \|\bar{A}_{hi}\|, \qquad \bar{C} = \|\bar{C}_{hi}\| \tag{59}$$

Also, according to (35), (21) and (22), we can set

$$\phi = N/n\eta^2 \tag{60}$$

The subscript of the correlation ratio can be dropped here and henceforth.

From equations (56) through (60), (13) becomes

$$\bar{x}\bar{A} = \eta^2\bar{x} \tag{61}$$

and (55) becomes

$$\bar{x}\bar{C} = \frac{1}{\eta^2}\bar{x} \tag{62}$$

If none of the g_j vanish, it is seen from Table 7 that all scale types of persons must occur so that none of the f_i vanish. Then, according to (14) and (57), $\bar{A}$ is nonsingular and Gramian. Hence, there are exactly $m + 1$ independent vector solutions $\bar{x}$ to (61), corresponding to $m + 1$ positive latent roots η^2. Every solution of (61) satisfies (62), except for the constant solution. From its order, $\bar{C}$ has at most $m + 1$ independent latent vectors; m of these are also latent vectors of $\bar{A}$, but its latent roots are the *reciprocals* of the corresponding ones of $\bar{A}$. The constant vector now has the elements $\bar{x}_i \equiv \sqrt{f_i}$; it has the root $\eta^2 = 1$ in (61) and the root $1/\eta^2 = 0$ in (62). Hence, every vector solution of (62) also satisfies (61), allowing for the difference in latent roots for the constant vector.

It is well known that the inverse of a symmetric matrix has the same latent vectors as the matrix, and that the latent roots of the inverse are the reciprocals of the corresponding roots of the matrix. $\bar{C}$ almost fulfills this condition with respect to $\bar{A}$. Of course, $\bar{C}$ is singular, since it has a zero root. However, if the zero root were replaced by a root of unity, then $\bar{C}$ would become nonsingular and equal to $\bar{A}^{-1}$. The replacing of the zero root by unity is accomplished by adding $\sqrt{f_h f_i/N}$ to $\bar{C}_{hi}$ to form the matrix R defined by

$$R_{hi} = \bar{C}_{hi} + \sqrt{f_h f_i/N}, \qquad (h, i = 0, 1, \cdots, m) \tag{63}$$

Then

$$R = \bar{A}^{-1} \tag{64}$$

Thus, we have the interesting result that the difference equations involve essentially the *inverse matrix* of the matrix in the original stationary equations for scores.

The inversion for weights. An exactly similar situation holds for the equations for weights. It is convenient to change notation slightly and to let

$$z_j = F_j y_j \qquad (j = 1, 2, \cdots, 2m) \tag{65}$$

Then equations (48), (46), and (50) can be written in the respective expanded forms:

$$z_1(1/f_0 + 1/f_1) - z_2/f_1 = \frac{\phi}{c_0} z_1 \tag{66}$$

$$-z_j/f_j + z_{j+1}(1/f_j + 1/f_{j+1}) - z_{j+2}/f_{j+1} = \frac{\phi}{c_j} z_{j+1} \tag{67}$$

$$(j = 1, 2, \cdots, m - 2)$$

$$-z_{m-1}/f_{m-1} + z_m(1/f_{m-1} + 1/f_m) = \frac{\phi}{c_{m-1}} z_m \tag{68}$$

Let F_{jk} $(j, k = 1, 2, \cdots, m)$ be the elements of the symmetric matrix F defined as

$$F = \left\| \begin{matrix} \frac{1}{f_0} + \frac{1}{f_1} & \frac{-1}{f_1} & & & & \\ \frac{-1}{f_1} & \frac{1}{f_1} + \frac{1}{f_2} & \frac{-1}{f_2} & & & \\ & \frac{-1}{f_2} & \frac{1}{f_2} + \frac{1}{f_3} & \frac{-1}{f_3} & & \\ & & \cdots\cdots & & & \\ & & & \frac{-1}{f_{m-2}} & \frac{1}{f_{m-2}} + \frac{1}{f_{m-1}} & \frac{-1}{f_{m-1}} \\ & & & & \frac{-1}{f_{m-1}} & \frac{1}{f_{m-1}} + \frac{1}{f_m} \end{matrix} \right\| \tag{69}$$

Notice how F differs in form from C in (54) in its first and last elements. Equations (66), (67), and (68) combine into the single matric equation

$$\sum_{j=1}^{m} z_j F_{jk} = \frac{\phi}{c_{k-1}} z_k \qquad (k = 1, 2, \cdots, m) \tag{70}$$

Equation (70) is not directly comparable to (16) because the square matrices involved are of orders m and $2m$ respectively. To reduce (16) in terms of a matrix of order m, we first verify from (17), (3), and (9) that

$$B_{j,\,k+m} = \frac{F_j}{n} - B_{jk} \qquad \begin{pmatrix} j = 1, 2, \cdots, 2m \\ k = 1, 2, \cdots, m \end{pmatrix} \tag{71}$$

Also, from (24) and (65),

$$z_{k+m} = -z_k \qquad (k = 1, 2, \cdots, m) \tag{72}$$

Hence, (16) becomes, from (65), (71), (72), (5), (11), and (34),

$$\sum_{j=1}^{m} z_j(NB_{jk} - F_jF_k/n)/c_{j-1} = \eta^2 z_k \qquad (k = 1, 2, \cdots, m) \tag{73}$$

Let:

$$\bar{z}_j = z_j/\sqrt{c_{j-1}} \tag{74}$$

$$\bar{B}_{jk} = (NB_{jk} - F_jF_k/n)/\sqrt{c_{j-1}c_{k-1}} \qquad (j, k = 1, 2, \cdots, m) \tag{75}$$

$$\bar{F}_{jk} = \frac{n}{N} F_{jk}\sqrt{c_{j-1}c_{k-1}} \qquad (j, k = 1, 2, \cdots, m) \tag{76}$$

and let

$$\bar{z} = ||z_j||, \qquad \bar{B} = ||B_{jk}||, \qquad \bar{F} = ||F_{jk}|| \tag{77}$$

Then (73) and (70) become, respectively, remembering (60),

$$\bar{z}\bar{B} = \eta^2\bar{z} \tag{78}$$

$$\bar{z}\bar{F} = \frac{1}{\eta^2}\bar{z} \tag{79}$$

It is easily seen that $\bar{F}$ is precisely the inverse of $\bar{B}$. In contrast to $\bar{C}$, $\bar{B}$ is nonsingular as it stands, for it has m positive latent roots and is of order m. Every solution of (16) which satisfies (24) must also satisfy (78). There is one such solution corresponding to each nontrivial latent root of $\bar{A}$ because of (22). Since $\bar{A}$ has a latent root for each of the m nonconstant latent vectors, all of which roots are positive, $\bar{B}$ has these same latent roots and is nonsingular.

Then, from (78) and (79), the m latent roots of F are the reciprocals of those of $\bar{B}$, so that

$$\bar{F} = \bar{B}^{-1} \tag{80}$$

The latent roots are all distinct. It is of considerable interest to know whether or not the latent roots are distinct. As is well known, if any roots of a symmetric matrix are distinct, then the corresponding latent vectors are *orthogonal* to each other. We shall establish the distinctness by showing that any two latent vectors having equal roots must be the same vector, which will contradict the well-known theorem that any symmetric matrix has as many linearly independent latent vectors as its order, regardless of the nature of the roots.

Suppose that u_i and v_i ($i = 0, 1, \cdots, m$) are vector solutions to (51), (52), and (53), with the respective roots ϕ_1 and ϕ_2. According to (51),

$$c_0(u_0 - u_1) = \phi_1 f_0 u_0 \tag{81}$$

$$c_0(v_0 - v_1) = \phi_2 f_0 v_0 \tag{82}$$

Multiply both members of (81) by v_0, and of (82) by u_0, and subtract the resulting second equation from the first:

$$c_0(u_0 v_1 - u_1 v_0) = (\phi_1 - \phi_2) f_0 u_0 v_0 \tag{83}$$

Hence, if $\phi_1 = \phi_2$, then the left member vanishes:

$$\begin{vmatrix} u_0 & u_1 \\ v_0 & v_1 \end{vmatrix} = 0 \tag{84}$$

From (52), we have

$$-u_i c_i + u_{i+1}(c_i + c_{i+1}) - u_{i+2} c_{i+1} = \phi_1 f_{i+1} u_{i+1} \qquad (i = 0, 1, \cdots, m-2) \tag{85}$$

$$-v_i c_i + v_{i+1}(c_i + c_{i+1}) - v_{i+2} c_{i+1} = \phi_2 f_{i+1} v_{i+1} \qquad (i = 0, 1, \cdots, m-2) \tag{86}$$

Multiplying both members of (85) by v_{i+1}, and of (86) by u_{i+1}, and then subtracting, yields

$$\begin{aligned} c_i(u_{i+1} v_i - u_i v_{i+1}) + c_{i+1}(u_{i+1} v_{i+2} - u_{i+2} v_{i+1}) \\ = (\phi_1 - \phi_2) f_{i+1} u_{i+1} v_{i+1} \qquad (i = 0, 1, \cdots, m-2) \end{aligned} \tag{87}$$

If $\phi_1 = \phi_2$, then the right member of (87) vanishes. Also, for $i = 0$, the first term on the left vanishes according to (84). Hence, the second term on the left must also vanish for $i = 0$, which implies

$$\begin{vmatrix} u_1 & u_2 \\ v_1 & v_2 \end{vmatrix} = 0 \tag{88}$$

From (88), the first term in (87) vanishes for $i = 1$, so that the second term must also vanish for $i = 1$. Continuing in this fashion, we see that, if $\phi_1 = \phi_2$, then

$$\begin{vmatrix} u_i & u_{i+1} \\ v_i & v_{i+1} \end{vmatrix} = 0, \qquad (i = 0, 1, \cdots, m - 1) \tag{89}$$

That (89) holds for $i = m$ could also have been seen from (53), according to which:

$$c_{m-1}(u_m - u_{m-1}) = \phi_1 f_m u_m \tag{90}$$

$$c_{m-1}(v_m - v_{m-1}) = \phi_2 f_m v_m \tag{91}$$

Multiplying both members of (90) by v_m, and of (91) by u_m, and subtracting, yields

$$c_{m-1}(u_m v_{m-1} - u_{m-1} v_m) = (\phi_1 - \phi_2) f_m u_m v_m \tag{92}$$

Hence, if $\phi_1 = \phi_2$, the left member must vanish, or (89) holds for $i = m$.

It follows from (89) that the matrix

$$\begin{Vmatrix} u_0 & u_1 & \cdots & u_m \\ v_0 & v_1 & \cdots & v_m \end{Vmatrix}$$

is of rank one, which is another way of saying that the u_i and v_i are proportional to each other and define the same (normalized) latent vector.

But this contradicts the fact that to two latent roots, equal or not, must correspond two linearly independent vectors. Hence, it must be that $\phi_1 \neq \phi_2$, or all the latent roots are distinct. This includes the trivial as well as the nontrivial roots.

Since the roots are all distinct, it follows that all the latent vectors

of $\bar{A}$ are orthogonal to each other, and similarly for the latent vectors of $\bar{B}$. According to (56), this implies that the score vectors are orthogonal with respect to the f_i:

$$\sum_{i=0} f_i u_i v_i = 0 \tag{93}$$

According to (74), (65), and (34), the weight vectors are orthogonal with respect to the $g_j F_j / F_{j+m}$:

$$\sum_{j=1} \frac{g_j F_j w_j y_j}{F_{j+m}} = 0 \tag{94}$$

where w_j and y_j are any two different solutions of (16).

Sum of the nontrivial roots is unity. It is well known that the sum of the latent roots of a matrix is equal to the sum of the elements in its principal diagonal. Hence to obtain the sum of the latent roots of $\bar{A}$ we need only to evaluate the sum of $\bar{A}_{ii}$ $(i = 0, 1, \cdots, m)$. From (57) and (14), noticing that $e^2_{ij} \equiv e_{ij}$,

$$\bar{A}_{ii} = \frac{f_i}{n} \sum_{j=1}^{2m} \frac{g_j e_{ij}}{F_j} \qquad (i = 0, 1, \cdots, m) \tag{95}$$

Summing both members over i, and using (9), we obtain

$$\sum_{i=0}^{m} \bar{A}_{ii} = \frac{1}{n} \sum_{j=1}^{2m} g_j \tag{96}$$

Hence, from (5) and (6),

$$\sum_{i=0} \bar{A}_{ii} = 2 \tag{97}$$

The trivial root of unity is included in the sum of all the latent roots implied by (97). Therefore, the sum of the nontrivial roots is unity. This means that the squares of all m correlation ratios add up to unity, or, the total variance, taken as unity, has been divided into m additive parts.

The alternating signs of bending points. The difference equations provide an easier means of studying the nature of the principal components than do the matric equations. Through them, we can now show the oscillatory nature of the latent vectors, both for scores and for weights. We begin with the scores, and shall show first how each bending point in the curve of a principal component, like those

portrayed in Figure 1, must be of the opposite sign from its predecessor.

Suppose that x_a is a *convex* bending point, defined by the conditions

(98) $$x_a \geqq x_{a-1}\ ,\ x_a > x_{a+1}$$

This implies

(99) $$\Delta_a x_{a-1} \geqq 0, \qquad \Delta_a x_a < 0$$

so that

(100) $$c_a \Delta_a x_a - c_{a-1} \Delta_a x_{a-1} < 0$$

Hence, from (100) and (38), setting $i + 1 = a$,

(101) $$x_a > 0$$

In exactly an analogous manner, if x_b is a *concave* bending point defined by

(102) $$x_b \leqq x_{b-1}\ ,\ x_b < x_{b+1}$$

it follows that

(103) $$x_b < 0$$

Since bending points must be alternately concave and convex, it follows that *successive bending points alternate in sign.*

In particular, (101) and (103) show that *no bending point can vanish.*

We shall further show that the first internal bending point must be opposite in sign to the first point, x_0, and that the last internal bending point must be opposite in sign to the last point, x_m.

First, we notice that x_0 can never vanish, for then x_1 would vanish according to (51), and then the remaining x_i would all vanish according to (52). Since latent vectors are determined only up to a constant of proportionality, there is no loss of generality if we set

(104) $$x_0 = 1$$

Then (51) becomes

(105) $$c_0(1 - x_1) = \phi f_0$$

Since the right member of (105) is positive, it must be that

$$x_1 < 1 \tag{106}$$

This means that, if the metric of (104) is used for all the principal components, each of their curves starts out with a downward slope, so that the first bending point (if there is any) must be concave. But all concave bending points are negative, according to (103); and x_0 is positive, according to (104). Therefore, the first bending point must be opposite in sign from x_0.

In an analogous manner, we find that x_m can never vanish, and it must be opposite in sign to the last turning point.

Therefore, for any nonconstant principal component, if the x_i are plotted as a function of i as in Figure 1, and if the points are connected by a polygon, then the polygon must cross the i-axis between each two successive turning points, counting the first and last points also as turning points.

Proof of the number of oscillations. Let us consider again two solutions, u_i and v_i, with their respective—and distinct—roots ϕ_1 and ϕ_2. Let us suppose that ϕ_1 is the larger of the two roots, and let δ be the difference, so that

$$\delta = \phi_1 - \phi_2 > 0 \tag{107}$$

We want to prove that the curve of the u_i, plotted as a function of i, bends at least once more than does that of the v_i. That is, we want to show that if one set of component scores has a *smaller* correlation ratio than the other (and hence a larger ϕ), then it also has at least one more oscillation in its curvilinear relation to i. First we shall show that the u_i and v_i cannot change in sign equally often, and hence cannot have the same number of turning points. Then it will be seen that[17] the u_i bend more often than do the v_i.

Let t_i be defined by

$$t_i = c_i(u_i v_{i+1} - u_{i+1} v_i), \qquad (i = 0, 1, \cdots, m - 1) \tag{108}$$

Then (83), (87) and (92) can be written, respectively, as

$$t_0 = \delta f_0 u_0 v_0 \tag{109}$$

[17] In this paragraph we tacitly assume that neither u nor v is the constant vector, so that their latent roots correspond to actual correlation ratios. But the constant solution has no bends, and neither does the metric vector, so these two solutions "oscillate" equally often. However, the constant solution has no sign changes, whereas the metric vector has one sign change. The proof that follows includes the constant vector.

$$t_{i+1} - t_i = \delta f_{i+1} u_{i+1} v_{i+1} \qquad (i = 0, 1, \cdots, m-2) \tag{110}$$

$$-t_{m-1} = \delta f_m u_m v_m \tag{111}$$

Suppose that the u_i and v_i change sign equally often. Then without loss of generality, since endpoints cannot vanish according to the preceding section, we can set:

$$u_0 v_0 > 0 \tag{112}$$

$$u_m v_m > 0 \tag{113}$$

Let a be the first value of i for which $u_i v_i < 0$, so that

$$u_i v_i \geqq 0, \qquad (i = 1, 2, \cdots, a-1) \tag{114}$$

$$u_a v_a < 0 \tag{115}$$

That there must be an $a < m$ satisfying (114) and (115) follows from the fact that the u_i and v_i are orthogonal as in (93).

From (109), (112) and (110) it follows that $t_{a-1} > 0$, whence, remembering (108),

$$u_{a-1} v_a > u_a v_{a-1} \tag{116}$$

Dividing both members of (116) by $u_a v_a$, remembering (115), shows that

$$\frac{u_{a-1}}{u_a} < \frac{v_{a-1}}{v_a} \tag{117}$$

Therefore, if v_a and v_{a-1} are opposite in sign, then also u_a and u_{a-1} are opposite in sign; but this contradicts (115) and (114) when $i = a - 1$. Hence, according to (115) and (114), it must be that

$$u_{a-1} u_a \leqq 0 \ , \ v_{a-1} v_a \geqq 0 \tag{118}$$

At most, only one of the equalities in (118) can hold, for according to (117), u_{a-1} and v_{a-1} cannot vanish simultaneously. If $v_{a-1} = 0$, then there is a change in sign between u_{a-1} and u_a. If $u_{a-1} = 0$, then v_{a-1} and v_a have the same sign; but u_{a-2} and u_a have opposite signs as is easily seen from (85) when $i + 1 = a - 1$; hence also v_{a-2} and v_a must have the same sign to conform to (115) and to (114) when $i = a - 2$. If neither u_{a-1} nor v_{a-1} vanishes,

then (118) shows a change in sign between u_{a-1} and u_a but no change between v_{a-1} and v_a.

In any event, then, whether or not $u_{a-1}v_{a-1} = 0$, the u_i have one more sign change than the v_i in the interval $0 \leqq i \leqq a$.

Let b be the first value of i after a for which $u_iv_i > 0$, so that

$$(119) \qquad u_iv_i \leqq 0, \qquad (i = a+1, a+2, \cdots, b-1)$$

$$(120) \qquad u_bv_b > 0$$

Two possibilities need be considered. Either $t_{b-1} < 0$ or $t_{b-1} \geqq 0$.

Suppose that $t_{b-1} < 0$. Then from (108)

$$(121) \qquad u_{b-1}v_b < u_bv_{b-1} \qquad (t_{b-1} < 0)$$

Dividing both members of (121) by u_bv_b shows that

$$(122) \qquad \frac{u_{b-1}}{u_b} < \frac{v_{b-1}}{v_b} \qquad (t_{b-1} < 0)$$

Reasoning as with (117), but using (119) and (120) instead of (114) and (115), shows that the u_i must have one more sign change than the v_i in the interval $a + 1 \leqq i \leqq b$.

If $t_{b-1} \geqq 0$, then instead of (122) we have

$$(123) \qquad \frac{u_{b-1}}{u_b} \geqq \frac{v_{b-1}}{v_b} \qquad (t_{b-1} \geqq 0)$$

whence the v_i can possibly have one more change than the u_i in the interval $a + 1 \leqq i \leqq b$, so that the u_i and v_i could have the same number of sign changes in the interval $0 \leqq i \leqq b$. In this case, it must be that $b < m$ in order that (111) and (113) be satisfied. Then we can go on to c, the first value of i after b for which $u_iv_i < 0$:

$$(124) \qquad u_iv_i \geqq 0, \qquad (i = b+1, b+2, \cdots, c-1)$$

$$(125) \qquad u_cv_c < 0$$

That there must be a value of $i > b$ for which $u_iv_i < 0$ follows from the fact that $t_i > 0$ for all $i > b$ such that $u_hv_h \geqq 0$ $(h = b + 1, \cdots, i)$, according to (110), (120) and the hypothesis that $t_{b-1} \geqq 0$, yet (111) and (113) must hold. In particular,

$$(126) \qquad t_{c-1} > 0, \qquad (t_{b-1} \geqq 0)$$

Hence, from (126), (108), and (125),

$$\frac{u_{c-1}}{u_c} < \frac{v_{c-1}}{v_c}, \qquad (t_{b-1} \geqq 0) \tag{127}$$

Equations (124), (125), and (127) are precisely of the same form as (114), (115), and (117) respectively. Hence, the u_i must have one more change of sign than the v_i in the interval $b + 1 \leqq i \leqq c$.

Hence, whether $t_{b-1} < 0$ or $t_{b-1} \geqq 0$, the u_i must change in sign more often than the v_i in the interval $0 \leqq i \leqq c$. Continuing in this fashion, it follows that the u_i change in sign more often than the v_i in the interval $0 \leqq i \leqq m$, which contradicts the hypothesis that the u_i and v_i have the same number of sign changes. Therefore, *no two different latent vectors can have the same number of sign changes.*

There being but $m + 1$ elements in a latent vector, no such vector can have more than m changes in sign. But there are $m + 1$ latent vectors, no two of which can have the same number of sign changes. Hence, there must be a vector with *no* sign changes, a vector with *one* sign change, a vector with *two* sign changes, etc., and a vector with m sign changes. Furthermore, a review of the proof shows that a vector with a smaller correlation ratio than another has the larger number of sign changes.

The vector with no sign changes is the constant vector. The vector with one sign change also has no bends, and is the *metric* vector for the scale, being a monotonic function of the rank order. The vector with two sign changes must have one bend, and is identified as the intensity function. The vector with three sign changes must have two bends; that with four sign changes must have three bends; etc. The vectors with two or more bends are yet to be identified psychologically.[18]

The oscillations for weights. The weight components oscillate in a manner similar to the score components. The difference equation (46) is the same in form as (38), but the boundary conditions vary slightly, so the nature of the oscillations is not quite the same. Also, we have to consider all the $2m$ elements in a weight component, instead of the $m + 1$ elements in a score component.

How the weights oscillate can be seen from equation (30). The vector $\Delta_i x_i$ ($i = 0, 1, \cdots, m - 1$) must have one less change in sign than the vector x_i ($i = 0, 1, \cdots, m$). But from (30), the

[18] See footnote 2 above added in proof.

weight vector y_{i+1} $(i = 0, 1, \cdots, m-1)$ changes sign whenever $\Delta_i x_i$ does. Therefore, the first m elements of the weight component y_j $(j = 1, 2, \cdots, 2m)$ change in sign once less than do the elements of the latent score vector belonging to the same correlation ratio. The last m elements change in sign similarly, by virtue of (24).

Particular solutions: the classical orthogonal functions. Difference equation (38), with its boundary conditions (41) and (44), brings to mind the classical Sturm-Liouville boundary value problem, which has as particular solutions the Legendre polynomials, Hermite polynomials, Laguerre polynomials, etc. These classical functions are orthogonal, and oscillate much as do the solutions to (38). Actually, these classical orthogonal functions can be seen to be obtained in the limit from solutions to (38) as m becomes indefinitely large.

The Sturm-Liouville problem[19] concerns the solutions $y(x)$ to the differential equation

$$\frac{d}{dx}\left[c(x)\frac{dy}{dx}\right] + \phi f(x)y(x) = 0$$

over an interval $a \leqq x \leqq b$, where $c(x)$ and $f(x)$ are given non-negative functions and ϕ is a constant, with boundary conditions

$$c(a)y'(a) + \phi f(a)y(a) = 0$$

$$c(b)y'(b) + \phi f(b)y(b) = 0$$

It is clear that our difference equation problem is a *discrete* analogue to the Sturm-Liouville problem, which concerns continuous functions, so that it should not be surprising that the limits of our solutions should be solutions to the continuous case.

The general discrete case does not seem to have been noticed before, although special cases have been studied. Tchebychef[20] studied a particular discrete solution which in the limit becomes the Legendre polynomials. The Krawtchouk polynomials[21] are a solu-

[19] E. L. Ince, *Ordinary Differential Equations* (Longmans, Green and Co., Ltd., London, 1927), p. 223–253.

[20] "Sur une nouvelle série," *Oeuvres*, I, pp. 381–384; "Sur l'interpolation," *ibid.*, I, pp. 541–560; "Sur l'interpolation par la méthode des moindres carrés," *ibid.*, I, pp. 473–498; "Sur l'interpolation des valeurs équidistants," *ibid.*, II, pp. 219–242; "Sur une formule d'analyse," *ibid.*, I, pp. 701–702.

For the approach to the Legendre polynomials, see G. Szegö, *Orthogonal Polynomials* (Colloquium Publications, Vol. 23, American Mathematical Society, New York, 1939).

[21] See M. J. Gottlieb, "Concerning Some Polynomials Orthogonal on a Finite or Enumerable Set of Points," *American Journal of Mathematics*, Vol. 60, No. 2 (April 1938), pp. 453–458.

tion to (38) which in the limit yield the Hermite polynomials. Charlier developed a series of polynomials which can be seen to satisfy (38), and in the limit yield the Poisson-Charlier polynomials.[22] Gottlieb[23] developed a series of polynomials which also can be seen to satisfy (38), and which in the limit yield the Laguerre polynomials.

None of these treatments of particular solutions seems to have recognized the general case expressed by equation (38). Perhaps the reason for this is the fact that the second term in (38) involves $i + 1$ rather than i. For example, Tchebychef's case is that for which (in our notation) $f_i \equiv 1$, $g_j \equiv 1$, so that

$$c_i = (i + 1)(N - i - 1), \qquad (i = 0, 1, \cdots, n)$$

and $N = m + 1$. Expanding the first term of (38) with these values of c_i we can rewrite (38) as

$$(i + 2)\,(N - i - 2)\,\Delta_i{}^2 x_i + (N - 2i - 3)\,\Delta_i x_i + \phi x_{i+1} = 0$$

Tchebychef, however, expressed the difference equation in the form[24]

$$(i + 2)\,(N - i - 2)\,\Delta_i{}^2 x_i + (N - 2i - 3 + \phi)\,\Delta_i\, x_i + \phi x_i = 0$$

Using this form as a starting point, it is not too obvious that a generalization like (38) is to be obtained; the temptation is strong to keep x_i in the last term on the left and not to use x_{i+1}. It seems fortunate that the problem of scale analysis led to the general equation (38) in a natural fashion.

The numerical example used in the first half of this chapter is Tchebychef's solution for $m = 5$ $(N = 6)$, since Tchebychef's case, in our setup, is that of a rectangular distribution for types of people and for types of items. Nonconstant values for the f_i and/or the g_j yield different solutions, including the various sets of orthogonal polynomials previously referred to.

In general, however, the orthogonal functions obtained from nonconstant frequencies need not be polynomials. For example, a particularly simple solution is for a case which is well known in the

[22] *Ibid.*
[23] *Ibid.*
[24] "Sur l'interpolation," *op. cit.*, p. 557.

theory of difference equations. If $f_i \equiv 1$ and $c_i \equiv 1$ (that is, $g_{i+1} \equiv F_{i+1}F_{i+m+1}$), then (38) and its boundary conditions become:

$$\Delta_i^2 x_i + \phi x_{i+1} = 0, \qquad (i = 0, 1, \cdots, m - 2)$$

$$\Delta_0 x_0 + \phi \dot{x}_0 = 0$$

$$\Delta_m x_{m-1} - \phi x_m = 0$$

These equations traditionally have been written in expanded form, rather than in the above form, again perhaps because of lack of recognition of a possibility of a generalization like (38). The solution consists simply of cosines.[25]

Reproducibility from the components. Just as the principal component scores and weights are derived from the item responses, just so the responses can be reproduced from the components. This reproducibility, in fact, holds whether or not the e_{ij} form a scale pattern, as we shall see from (20).

In order to discuss all the components simultaneously, it is necessary to modify previous notation. Let $\eta_0 = 1$, and let $\eta_1, \eta_2, \cdots, \eta_m$ be the respective m correlation ratios for the m principal component vectors satisfying (13). (If desired, it can be assumed that $\eta_1 > \eta_2 > \cdots > \eta_m$ so that the latent vectors have the order: constant vector, metric vector, intensity vector, etc.) Let x_{ai} be the ith element in the vector corresponding to η_a $(a, i = 0, 1, \cdots, m)$. Similarly, let y_{aj} be the jth element in the latent vector satisfying (16) with $\eta_y = \eta_a$ $(a = 0, 1, \cdots, m;\ j = 1, 2, \cdots, 2m)$. Then (20) can be written as

$$(128) \qquad \sum_{i=0}^{m} x_{ai} f_i e_{ij} = \mu_a y_{aj} F_j \qquad \begin{pmatrix} a = 0, 1, \cdots, m \\ j = 1, 2, \cdots, 2m \end{pmatrix}$$

According to (93), if $a \neq b$, then the x_{ai} and x_{bi} are orthogonal with respect to the f_i. Since the vectors are determined only up to a constant of proportionality, we can normalize them so that

$$(129) \qquad \sum_{i=0}^{m} f_i x_{ai}^2 = 1, \qquad (a = 0, 1, \cdots, m)$$

Then, from (93) and (129),

[25] Theodore V. Karman and Maurice A. Biot, *Mathematical Methods in Engineering* (McGraw-Hill Book Company, Inc., New York, 1940), pp. 453ff.

(130) $$\sum_{i=0}^{m} f_i x_{ai} x_{bi} = \delta_{ab} \qquad (a, b = 0, 1, \cdots, m)$$

where δ_{ab} is Kronecker's delta. Equations (130) imply that the matrix $\| x_{ai}\sqrt{f_i} \|$ is orthogonal, so that also its transpose is orthogonal:

(131) $$\sum_{a=0}^{m} x_{ah} x_{ai} \sqrt{f_h f_i} = \delta_{hi} \qquad (h, i = 0, 1, \cdots, m)$$

Multiply both members of (128) by $x_{ah}\sqrt{f_h}$ and sum over a. Using (131), and rewriting subscripts, we see that

(132) $$e_{ij} = \sum_{a=0}^{m} x_{ai} w_{aj}$$

where

(133) $$w_{aj} = \mu_a F_j y_{aj} \qquad \begin{pmatrix} a = 0, 1, \cdots, m \\ j = 1, 2, \cdots, 2m \end{pmatrix}$$

Equations (132) show how the response of the ith type of person to the jth type of item is a linear function of his principal component scores. Each type of category has its own set of weights w_{aj}, and each type of person has his own set of scores. The responses to the items can be reproduced algebraically from the weights and scores according to (132), just as well as they can be from the scale pattern defined in (2). According to (2), the item responses are a simple function of the rank order of people; according to (132), the item responses are a linear function of m principal components. Here is a paradox, that in one case reproducibility is from but *one* variable whereas in the other case reproducibility is from m variables. The paradox is resolved by remembering that the principal components in turn are perfect functions of the rank order.

A resolution into components like (132) is obtained whether or not e_{ij} is a scale pattern. The difference between the scalable and nonscalable cases is that, in the latter, the principal components are *not* in general perfect functions of a rank order.

External prediction. An important property of a scale is the ease with which the predictive power of its items can be used for any arbitrary outside variable. Just as the scale ranks can perfectly reproduce each item in the scale, just so are the scale ranks perfectly

efficient for any external purpose. It is easily seen that *the conditional distribution of any variable on the items in a scale is precisely the same as its conditional distribution with the scale ranks.* Indeed, this theorem borders on being a tautology. A conditional distribution with the items themselves is a set of conditional probabilities, one for each combination of item responses. But each combination of item responses has but one scale rank. Conversely, each scale rank corresponds to but one combination of responses. Hence, the set of conditional probabilities for scale ranks is identical with that for item responses. This means that the multiple regression of any outside variable on the items is equivalent to its simple regression on the scale ranks. The same predictions and the same errors of predictions are obtained either way.

The resolution into components expressed by (132) adds an interesting view to the problem of external prediction. In the case of *linear* regression of an outside variable on the items, the prediction for the ith type of person is of the form

$$p_i = \sum_{j=1}^{2m} b_j e_{ij} \qquad (i = 0, 1, \cdots, m) \tag{134}$$

where b_j are regression weights. According to (132), then,

$$p_i = \sum_{a=0}^{m} x_{ai} w_a' \qquad (i = 0, 1, \cdots, m) \tag{135}$$

where

$$w_a' = \sum_{j=1} b_j w_{aj} \qquad (a = 0, 1, \cdots, m) \tag{136}$$

Since the x_{ai} are orthogonal (with respect to the f_i), each w_a' is proportional to the zero-order correlation of the outside variable with x_{ai}. If p_i is a linear function of x_{1i}, then $w_a' = 0$ for $a = 2, 3, \cdots, m$. That is, *if the regression is a linear function of the metric component, then the regression weights for the intensity and further components are all zero.* The intensity and further components get nonzero weights only when the regression on the metric component is curvilinear.

CHAPTER 10

THE LOGICAL AND MATHEMATICAL FOUNDATION OF LATENT STRUCTURE ANALYSIS[1]

Introduction

CHAPTERS 10 and 11 describe the latent structure approach to the treatment of itemized tests. Each chapter is divided into three sections. The first section of Chapter 10 develops the general theory of "trace lines" and introduces their "equivalent latent structures." This section is written primarily for the nonmathematician. In the second section the general algebraic relationships between the latent structure and manifest (empirical) test data are derived. The third section shows in a number of specific cases how the latent parameters can be obtained from manifest data if certain conditions of reducibility are satisfied.

In Chapter 11 numerical examples are introduced using Research Branch data. Section IV deals first with a latent dichotomy and discusses in detail problems of testing empirical data against a presumed latent structure. Next the techniques are applied to a special case of a latent trichotomy. The latter is shown to be closely related to a Thurstone attitude scale. In the second section data are investigated which correspond to a "social distance" test and to Guttman's scalogram pattern. There also an example is given which permits discussion of the empirical handling of several latent characteristics. In the third section the relation of the latent structure theory to the general problem of scaling is further

---

[1] By Paul F. Lazarsfeld. The investigations presented in this and the following chapter were developed in connection with work the writer did as a consultant to the Research Branch. Some of the general ideas were discussed between the writer and Professor S. A. Stouffer in connection with earlier joint research efforts. One cannot overestimate the influence which these many years of intellectual cooperation have had on the present text. No further effort will be made to trace in detail all the help and advice Dr. Stouffer has given during the writing of Chapters 10 and 11.

discussed. This provides an occasion to consider a number of other measurement problems as they have appeared in the literature on itemized tests.

The six sections in Chapters 10 and 11 are numbered consecutively. An equation numbered 4.7 is therefore equation 7 in section 4.

SECTION I

GENERAL BACKGROUND

1. THE NOTION OF TRACE LINES

The practice of testing can be roughly described as follows: The investigator assumes that a one-dimensional continuum exists, such as soldier morale, anti-Semitism or intelligence. People are assumed to be arranged on this continuum in an unknown way. Then, persons are exposed to a series of test items. They can answer questions with yes-no, true-false, agree-disagree, etc.; or they can be observed as to whether they do or don't do certain things; or they can be given arithmetic problems to see whether they can or cannot solve them. Any verbal or other behavior which permits a dichotomous classification of people will hereafter be called a test item.

The investigator chooses items which seem to him relevant to the problem. In a test of soldier morale he will not usually include a problem of integral calculus, and in an intelligence test he will not include a question as to whether the respondents like the food they are getting in the Army. (The reader will remember that we are giving a description of prevailing practice and not discussing how tests should be or would be constructed.)

After the test items are selected, they are given to a sample of people. The purpose of giving the test is obviously the following one: From the distribution of "answers" to the items we can, it is hoped, make inferences as to the position of the people on the assumed continuum. While we might not be able to say metrically where each respondent belongs on the continuum, we want at least to rank people according to their position there. The scoring system developed for a test always implies certain assumptions, however vague, about the relations between performance on the items and the hypothetical continuum. Because the "response patterns" of the people are obtained from actual experiments, we shall call them *manifest data;* all the information inferred as to the nature of the continuum or the position of people thereon we shall call *latent.*

Nothing more is implied in this terminology than the distinction between information obtained by direct observation and information obtained by using additional assumptions and/or by making inferences from the original data.

We shall now translate the foregoing description into more precise terms. The continuum we shall call x and represent it on the x-axis of a coordinate system. Nothing will be presumed about this continuum for the time being other than its existence, and even that will be made subject to empirical inquiry later.

Let us look at specific test items. Suppose in connection with studying the effect of an indoctrination film we want to test the "ethnocentrism" of American soldiers. By this we mean vaguely the extent to which they think that only Americans are of value and that the actions and ideals of the war allies are of little value or importance for the war effort. Suppose the following three items have been proposed for inclusion in the test:

1. I believe that our European allies are much superior to us in strategy and fighting morale. Yes ______ No ______

2. The majority of all equipment used by all the allies comes from American lend lease shipment. True ______ False ______

3. Neither we nor our allies could win the war if we didn't have each other's help. Agree ______ Disagree ______

All three items were included in the test because they are related and partially express ethnocentrism the way we have conceived it to begin with. This fact can be formulated more precisely in the following way: We assume that the probability y, that respondents check the first alternative in the ith question, is a function $f_i(x)$ of their position on the continuum x, their degree of ethnocentrism. This formulation can be graphically represented in Figure 1. The graphical picture of our functions $f_i(x)$ we shall call the *trace line* of item i.

The way these trace lines have been drawn can be expressed verbally as follows: According to Figure 1, item 1 is likely to be considered true only by people with very low ethnocentrism. Among persons with at least some ethnocentrism, hardly anyone will consider the statement true. Item 2 is more likely to be answered "yes" the more ethnocentric a respondent is. There is no special break in trace line 2: the probability y goes uniformly up with x, and the trace line can be expressed by the equation $y = a + bx$.

The third item is the typical "middle item" of any attitude test. Its trace line goes down toward the two extremes. The probability that the statement will be agreed with is highest in the middle range of the continuum x. Trace lines of this type obviously can vary greatly in the steepness of their peak.

In this discussion we shall deal only with dichotomous questions. (If questions permit three or more qualitative answers, the data can be reduced to a set of dichotomies, as will be shown at the end of Chapter 11.) No limitation is set on the kind of dichotomies which can be used. It would not change the theory, for instance, if

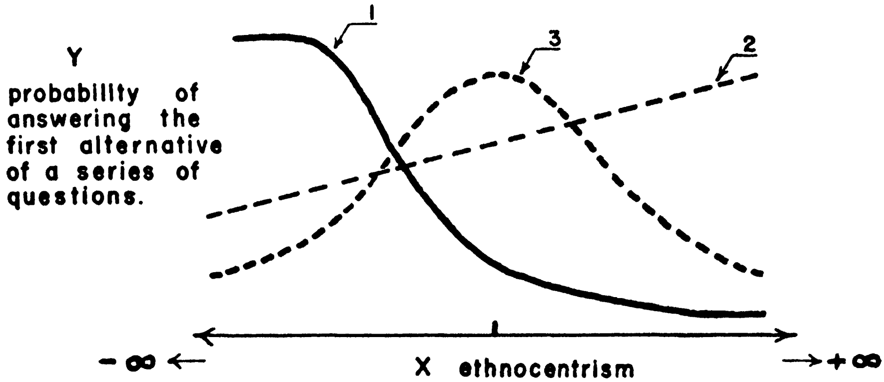


Figure 1. Trace lines of 3 items in a test of "ethnocentrism."

some of the items were observations to the effect that each person did or did not perform a certain act or own a certain object. It also is a matter of indifference which of the two answers is plotted in a trace line as long as the same practice is maintained all through the analysis. The answer which is used to formulate the trace line will be called the "positive answer." In this sense it can then very well be that we decide in a specific context to use the answer "no" as the "positive" alternative. The term "positive answer" will refer hereafter to the answer which is used to compute probabilities.

It is, of course, not assumed at this point that we actually know the trace lines for items. The contention is only that this whole presentation is an acceptable formalization of the reasoning which goes on when an item is included in a test. In practice, most of the time, we merely have a vague feeling that the trace line goes up or down from left to right or has its peak in the middle. (It will be shown later, however, that it is possible to make a good guess in advance as to whether the trace line "breaks" or whether it has a

linear shape.) Moreover, we have at this point discussed only the geometric shape of the trace lines and not their position up and down on the y-axis.

The gist of the discussion up to this point is as follows:

a. Items are included in tests because they are supposed to have some kind of relationship to the underlying continuum which is to be measured.

b. The reasoning of the test builder consists of a series of expectations. He expects people to be distributed over the underlying continuum; at each level of the continuum, he thinks, people will have a different probability of answering a test item "positively."

c. If the test builder could express his reasoning with perfect clarity and in full detail he would set down for each item a trace line. This would represent a mathematical function relating to each level of the continuum a definite probability for a positive answer.

d. In a concrete case, the actual reasoning of the test builder who chooses an item can usually be translated into his assuming some more or less precise properties of a hypothetical trace line.

2. AGGREGATES OF ITEMS WHICH FORM A "PURE" TEST

Although there is, of course, no way to verify such a set of assumptions directly, they represent what we think is a plausible description of the process of test building. But the description is made in mathematical terms. It leads to a model from which, as will be seen later, rather important mathematical inferences can be drawn. The inferences can be judged as to whether they are right or wrong. The original "formalization" can be judged only as to whether it is reasonable and useful.

The test builder not only makes assumptions about each separate item; he also judges the items in their interrelationships. This judgment enters in a variety of ways. Suppose a large number of items has been proposed for inclusion in a test of ethnocentrism. Someone is sure to ask how we know that this is a "unitary attitude." Maybe American soldiers are convinced of American technical superiority but believe in British cultural superiority. Should we have just one test or two tests? Or, suppose that among our questions there are a few which express practically the same idea with only minor changes in wording. Wouldn't a respondent who approves of such a particular idea get a much higher score than another respondent who prefers to express his ethnocentrism through another idea? One may also inquire how we know that an aggregate of items is a "pure" test of ethnocentrism; might it not measure ethnocentrism *and* something else?

All such considerations can, as will be seen, be reduced to an analysis of the *interrelations between the items.* Actually certain expectations as to these interrelationships are already implied in the preceding discussion. Let us take items 1 and 2 in our test of ethnocentrism. By and large, the items were picked because we assumed that soldiers high in ethnocentrism would think that most allied equipment comes from America and that European armies are not superior to American ones. (This expectation is expressed in the trace lines of items 1 and 2 in Figure 1.) Therefore, the answers to the two items will show a correlation—in this case a negative relationship. Respondents giving a positive reply to item 1 will be more likely to give a negative answer to item 2. We shall now call a *pure test* of a continuum x an aggregate of items which has the following properties: *All interrelationships between the items should be accounted for by the way in which each item alone is related to the latent continuum.*

Before giving this idea a mathematical formulation, some additional remarks are necessary. No requirement has been made that we should in practice always use pure tests. The notion is introduced only to make precise a number of distinctions which come up in practical work. Suppose ethnocentrism divides itself into two dimensions—cultural and technical. We should then be able to separate those items which are pure measures of the one and those which are pure measures of the other dimension; they would be those items which have flat horizontal trace lines with the one dimension they do not measure. Some of the items might well measure both dimensions. That is to say, their trace lines on both continua might be nonhorizontal. Nothing at this point permits us to decide whether it is a good idea to combine the three types of items in one test. The only purpose of the discussion is to find a way by which we can recognize what kind of aggregate of items we are treating. In a similar way, we would want our analysis to clarify what is meant by group factors in itemized tests, what is meant by biased questions, and so on. It will turn out that all of that can be achieved by comparing an actual aggregate of items with the special case of a pure test.

The crucial point is to obtain a precise formulation for the idea that the interrelationship between items should be completely explained by the existence of one underlying continuum. Scrutiny of research practice suggests that this be formulated as a negative statement. If a group of people has the same x-value—i.e., the

same degree of ethnocentricity—then, with ethnocentricity held constant, nothing else "holds the two questions together." Thus, for all people with any specific x-value, x_c, we have a fourfold table like the following:

		Item 1		
		Endorse	Do not endorse	
Item 2	Endorse			p_2
	Do not endorse			$1 - p_2$
		p_1	$1 - p_1$	1

In a pure test, there is no association between responses to the two items; therefore, the entry within the upper left-hand box, which we call p_{12}, is equal to $p_1 \times p_2$. In terms of trace lines, for any $x = x_c$, the probability of a joint positive answer $f_{12}(x)$ is the product of the independent probabilities for the two items under discussion:

$$f_{12}(x_c) = f_1(x_c) f_2(x_c)$$

But this should of course be true for every point of the continuum. So we come to the following *definition of a pure test:* It is an aggregate of items such that the trace line for a joint positive answer to several items $i, j, k, \cdots$ is the product of the trace lines of every item taken separately.

$$f_{ijk\ldots}(x) = f_i(x) f_j(x) f_k(x) \cdots \tag{1.1}$$

In actual experience, of course, it can easily happen that equation (1.1) is not fulfilled. There might be a little detail in the wording of two questions which "ties them together" beyond what can be accounted for by the underlying continuum. Thus for people who have the same x-value, we would find that $f_{12}(x)$ is greater or less than $f_1(x) f_2(x)$. In this case, we usually say that there has entered an additional factor, that we do not have a pure measurement of the continuum x. One should be sure not to think of equation (1.1) as describing manifest data obtained from actual test performances. If in the whole sample p_{12} per cent gives a positive answer to items

1 and 2, this value of p_{12} in general will *not* equal $p_1 \times p_2$. Thus is raised the decisive question of the relationship between trace lines and manifest response patterns in a concrete test situation.

But first it will be helpful to summarize the second step of the discussion which was just developed. This would be a sequence continuing the summary given at the end of paragraph 1.

e. It is possible to formulate mathematically what is meant if we say that an underlying continuum accounts for the interrelationship of two test items.

f. Such a formulation reduces to this idea: If people have the same position on the underlying x-continuum, then their answers to the two questions will be unrelated; the probability that they will answer two questions positively is then the product of the probabilities that they will answer each question alone positively.

g. This leads to the definition of a pure test: It is an aggregate of items such that the joint positive answers to any number of items have themselves a trace line which is the product of the original trace lines for the items viewed separately.

3. THE ACCOUNTING EQUATIONS

In our original discussion it was already implied that each respondent has an (unknown) degree of ethnocentricity (x). The total sample is therefore characterized by a distribution function $\phi(x)$ which gives for each small interval dx the number of people $\phi(x)dx$ whose score lies in this interval. We now can tell what proportion of respondents in the whole sample will give a positive reply to item i with trace line $f_i(x)$:

$$\int_{-\infty}^{\infty} \phi(x)\, f_i(x) dx = p_i \tag{1.2a}$$

The latter equation deserves a brief discussion for the nonmathematician. The symbol p_i stands for the proportion of people in the whole sample who check the first answer in each of our three items. This is usually called the *marginal* of the item. To indicate that p_i comes from actual observation, we shall sometimes call it the manifest marginal. In a small score range dx we can assume that the probability of a positive answer is $y = f(x)$ for all people. Therefore, within this range, the proportion of positive responses will be $\phi(x)\, f(x)dx$. For the whole interval, it will therefore be given by a definite integration over the whole continuum. This leads to

formula (1.2a) which, in essence, expresses the fact that the number of people who give a positive answer in the whole sample can be represented as a sum of all positive respondents taken from all strata which have a simi'ar degree of ethnocentricity.

Equation (1.2a) contains the whole answer to our problem. The manifest response patterns in a pure test could obviously be deduced from the trace lines if the population distribution were known. For we also know the trace lines for joint occurrences. So by reasoning identical with that just carried out, we get the following set of equations:

$$\int_{-\infty}^{+\infty} \phi(x)\, f_i(x)\, f_j(x)\, f_k(x) \cdots = p_{ijk\cdots} \tag{1.2b}$$

It is important to be clear as to the distinction between a trace line $y = f(x)$ and the population distribution $y = \phi(x)$. In both equations x represents the same continuum. But the meaning of y is different in the two equations. In the population distribution y represents the *proportion of all people out of the total sample* who have a certain degree of ethnocentricity. There is of course only one population distribution for the whole sample. But there are as many trace lines as there are items in the test. And for each trace line the value of y indicates *the proportion of people with a given degree of ethnocentricity who make a positive response to an item.*

Equations (1.2) tell how the manifest response frequencies, as they come from an empirical observation, are related to the hypothetical trace lines and $\phi(x)$. The frequencies on the right hand of the equations are known. The functions on the left side of the equations are unknown. The purpose of further discussion will be to decide what about $f(x)$ and $\phi(x)$ can be learned from these equations.

The term *accounting equations* is profitably used for (1.2). They express the postulate that the joint occurrences and therefore the interrelationships between the test items are accounted for by the existence of one underlying continuum. Inversely, they also describe the way in which the test "measures" this continuum. We can now add the following points to our summary:

h. A function $\phi(x)$ is introduced which describes the distribution of a sample of people over the x-continuum, i.e., the underlying characteristic which the test is devised to study.

i. It is possible to set up a series of accounting equations which show how the manifest response patterns on the one hand are related to the latent trace lines and the population distribution on the other. (Equations 1.2).

j. The problem is how one can, with the help of these accounting equations, learn the nature of the trace lines and the population distribution after an empirical test has provided the frequencies of all possible response patterns which are made by a sample of people.

4. THE USE OF POLYNOMIAL TRACE LINES

In their most general form, not much can be done with equations (1.2). As usual in such cases, we have to make certain simplifying assumptions if we want to carry out actual computation. A variety of such assumptions can be thought of. For the purpose on hand, it is most useful to assume that the trace lines are polynomials:

$$f_i(x) \equiv y = a_i + b_i x + c_i x^2 + \cdots$$

The use of polynomials has a number of advantages. It is well known that a very large variety of shapes can be approximated by polynomials of rather low degree. Even if the trace lines are rather complicated functions, they can most likely be developed into a Taylor's series and thus approximated by a polynomial. Furthermore, the accounting equations have some simple and useful properties if the trace lines are polynomials.[2] Consider, for instance, equation (1.2a). Assume for the sake of the example that the trace lines are all of second degree; then the manifest marginal of an item is related to its trace line in the following way:

$$(1.3a) \qquad p_i = a_i + b_i M^{(1)} + c_i M^{(2)}$$

In other words, the marginal is a linear aggregate of the moments $M^{(i)}$ of the distribution function. A similar result would be true for the joint occurrences. The relative frequency, for instance, with which respondents give a positive reply to two items would be given by the following formula:

$$(1.3b) \qquad p_{ij} = a_i a_j + (a_i b_j + a_j b_i)M^{(1)} + (a_i c_j + a_j c_i + b_i b_j)M^{(2)} + (b_i c_j + b_j c_i)M^{(3)} + c_i c_j M^{(4)}$$

[2] Professor L. J. Savage has pointed out to this writer that a more general assumption would be that all the trace lines are linearly dependent upon m of them. In such a more general theory the number m of independent trace lines would play the same role as the degree m of polynomial trace lines plays in the present text. For the purpose on hand, polynomial trace lines are sufficient to describe the relevant relationships between the items of a test and an assumed underlying continuum.

By combining the last two formulae, we find that the cross product, $[ij] = p_{ij} - p_i p_j$, between the two items takes on an especially simple form:

$$(1.3c)\quad [ij] = b_i b_j [M^{(2)} - M^{(1)2}] + c_i c_j (M^{(4)} - M^{(2)2}) + (b_i c_j + b_j c_i)(M^{(3)} - M^{(1)}M^{(2)})$$

By thinking in terms of polynomial trace lines, our task becomes narrowed. We have to solve equations of the type (1.3): on the left side there are relative frequencies known from empirical observation; on the right side, there are forms built from moments of the distribution function and from coefficients of the trace lines.

The concept of polynomial trace lines permits us to raise and answer an important question, namely: how should we decide on the degree of the polynomial which a given material requires? Should the trace lines be straight lines? Should they be quadratic parabolas as every high school student knows them from his study of conics? Should they be cubic parabolas or of even higher degree? It turns out that *the appropriate degree for the hypothetical trace lines can be derived from the manifest material.*

The general idea can be explained by the following example: Let us focus our attention on the joint occurrences of all pairs of two items which can be formed from a test. These joint occurrences can be set down in the form of a matrix. In the first row and in the first column of this matrix we put down the marginals of each item. Into each box we put one joint occurrence, from manifest observed data. The diagonal of the matrix is left open except for the left upper box where we put down the size of the sample, which will be 1, if we think of the marginals as proportions. The whole idea will be quite obvious from the following scheme:

$$\begin{matrix} 1 & p_1 & p_2 & p_3 & p_4 & \cdots \\ p_1 & \cdot & p_{12} & p_{13} & p_{14} & \cdots \\ p_2 & p_{21} & \cdot & p_{23} & p_{24} & \cdots \\ p_3 & p_{31} & p_{32} & \cdot & p_{34} & \cdots \end{matrix}$$

It can be proved that the rank of this matrix depends upon the degree of the polynomial trace lines, from which the joint frequencies are supposed to have generated. *If the degree of the trace lines is* m, *then the rank of this matrix will be* (m + 1). Any determinant in this matrix of order $(m + 2)$ will vanish.

The importance of this fact lies in its inversion. The matr x

which was just put down consists of empirical manifest data which we can inspect. If it has a rank of, say, 4—if all determinants of order higher than 4 vanish (after reasonable error allowances have been made)—then we can assume that the trace lines are of degree 3, that they are cubic parabolas. Actually, the conditions are more complicated because a similar rule also applies to the matrices which can be formed from the higher joint occurrences, for instance, of three items at a time. The full extent of these conditions will become clear in the second section of this chapter. What matters at the moment is this: *From an inspection of the manifest joint occurrences as they come up in a concrete test, we can infer the degree of the polynomial trace lines with which we have to work in a concrete case.*[3]

It will turn out later that the study of joint occurrence matrices is a crucial step in the approach to testing developed here. For this introduction, it is enough to keep the following two points in mind:

k. For a variety of reasons, it is helpful to assume that the trace lines are polynomials.

l. The degree of the polynomials is determined by the actual data in a given problem. This degree can be derived from a study of the joint occurrence matrices.

5. RELATION BETWEEN POLYNOMIAL TRACE LINES AND MANIFEST DATA

The first step then in any empirical investigation is to make a decision as to the degree of polynomial trace lines one will want to work with. Because of the fallibility of actual data, this will always be somewhat arbitrary, and we will come back to this point in the next chapter. At the moment, the question to be discussed is this: What do the joint frequencies tell us about the coefficients of the trace lines and the nature of the underlying distribution curve? To clarify the whole trend of thought, we shall discuss here the case of linear trace lines. We assume then that the joint frequencies of all order satisfy the necessary conditions and especially that the matrix of paired joint frequencies, which was exemplified above, has the rank 2. The following accounting equations then link the

[3] We are leaving out from this general discussion certain degenerate cases which are of considerable practical importance but which do not add anything to the general idea. Suppose, for instance, that all trace lines are cubic parabolas but all have one coefficient missing. Their formulae then would be $y = a + cx^2 + dx^3$ and the rank of the matrix would be 3 and not 4. The implications of such a possibility will be briefly referred to in the next section of this chapter.

manifest data to the coefficients of the linear trace lines and the moments of the distribution function (the use of the summation symbol is obvious):

(1.4a) $$p_i \;= a_i + b_i M^{(1)}$$

(1.4b) $$p_{ij} = a_i a_j + (a_i b_j + a_j b_i) M^{(1)} + b_i b_j M^{(2)}$$

(1.4c) $$p_{ijk} = a_i a_j a_k + \sum aab M^{(1)} + \sum abb M^{(2)} + b_i b_j b_k M^{(3)}$$

From (1.4a) and (1.4b) we deduce that:[4]

(1.5) $$[ij] = p_{ij} - p_i p_j = b_i b_j \{M^{(2)} - M^{(1)2}\} = b_i b_j \sigma^2$$

where $\sigma^2 = M^{(2)} - M^{(1)^2}$, the variance of the distribution curve $\phi(x)$ around its mean $M^{(1)} = M$.

If we take three items i, j, k we can deduce from (1.5) that

(1.6) $$b_i = \sqrt{\frac{[ij]\,[ik]}{[jk]\sigma^2}} = \frac{S_i}{\sigma}$$

This introduces the symbol $S_i = \sqrt{\dfrac{[ij]\,[ik]}{[jk]}}$

Solving (1.4a) for a_i, the equations of the trace lines then read

(1.7a) $$f_i(x) \equiv y_i = \left(p_i - \frac{MS_i}{\sigma}\right) + \frac{S_i}{\sigma} x$$

The coefficients of the trace lines are thus determined by the manifest data (p_i and S_i) and the two lowest moments of the distribution function (M and σ).

The implications of the result are clarified if one attempts to graph the trace lines on the assumption that the manifest data are given from an actual test. Obviously, this cannot be done, because we do not know the first two moments of the distribution curve.

[4] The symbol $[ij]$ to designate the "cross product" of a fourfold table will, in the next section, turn out to be the special case of a more general symbolism. There it will be applied to tables where the number of cases is not equal to 1. In the present preliminary discussion, where we deal with frequency *proportions* only, no distinction need be made between $[ij]$ pertaining to actual frequencies and $\frac{[ij]}{n^2}$, the standardized cross product. The distinction will appear, however, in the next section. A similar distinction will be necessary for the magnitude S_i to be introduced in equation (1.6).

Suppose, however, we graph one of the trace lines arbitrarily, say, through the origin and at an angle of 45 degrees with the x-axis. Then, relative to this one line, all the others are perfectly fixed. As can be seen from equation (1.7a) the first choice determines implicitly the two moments and thereafter the position and direction of all the other lines is determined, too. Another procedure would be to draw our (x,y) system fixing the average of the distribution curve at 0 and its standard deviation at 1. Then, again, the equations of the trace lines would be perfectly determined as follows:

$$f_i(x) \equiv y_i = p_i + S_i x \tag{1.7b}$$

Before pursuing further this point, another consequence of equations (1.4) should be noticed. From (1.4c) we can compute the third moment, which turns out to be:

$$M^{(3)} = \sigma^3 K + M(M^2 + 3\sigma^2) \tag{1.8}$$

$$\text{where } K = \frac{p_{123} - p_1 p_2 p_3}{S_1 S_2 S_3} - \left(\frac{p_1}{S_1} + \frac{p_2}{S_2} + \frac{p_3}{S_3}\right)$$

It is obvious that any further higher moments can be computed merely by adding additional accounting equations after (1.4c), starting with the joint occurrences for four items.

If the trace lines are of higher order, then the number of moments to be introduced rises rapidly. A manifest frequency p_{ijk} in the case of a cubic trace line, for example, would be represented by an equation of the type of (1.4c) where on the right (latent) side a ninth moment appears. It seems that then more than just the first two moments have to remain undetermined. The matter is still under investigation. For the time being we have to restrict ourselves to the following additions to our summary:

m. For polynomial trace lines the accounting equations link the manifest joint frequencies with the coefficients of the trace lines and the moments of the distribution function.
n. If the trace lines are linear the whole latent structure is determined but for a linear transformation.
o. This means that, for linear trace lines, we can compute the coefficients of the trace lines and any number of moments as long as we add further items to our test. Both coefficients and moments will be expressed as functions of the manifest data and the first two moments of $\phi(x)$.
p. In the case of higher degree polynomials more moments seem to remain undetermined. This general case has not yet been fully explored.

6. THE EQUIVALENT LATENT STRUCTURE

The results of the preceding paragraph show that we are not yet in the stage of performing actual computations. The introduction of polynomial trace lines led from the very general equations (1.2) to the much more specific equations (1.4); but, obviously, some further simplification has to be introduced if we want to get accounting equations where everything is determined. This step can be taken in a variety of directions. We could, for instance, establish how many unused independent accounting equations are left in the model after we have fixed the degree of the polynomial trace lines to be tested. Then we could introduce the most general distribution function with the given number of parameters. The accounting equations then could be solved for the coefficients of the trace lines and the parameters of $\phi(x)$.

In the present text one specific solution is followed up which corresponds to the assumption that $\phi(x)$ is a point distribution with $(m + 1)$ discrete points; if the total population is assumed to be 1, then its proportions concentrated at these $(m + 1)$ points are the m parameters of the distribution curve which are required for optimum results. Such a specification of $\phi(x)$ is tantamount to assuming *a partitioning of the whole population into a number of subclasses.* Each of these subclasses is assumed to be concentrated at a different point on the continuum. For intuitive illustration one might think of the way some writers talk about social classes. Implied in some social theories is the idea that society is stratified into a number of groups with rather distinct cleavages between them. One possible approximation to such an idea would be to assume that along some social stratification continuum the population has a discrete point distribution. The mathematics of such a model of segmentalization will be the main topic for the rest of this chapter.

Under the assumption of a discrete distribution there are no people outside a given number of points (class positions). The accounting equations, instead of being integrals, therefore become finite sums. The probabilities of answering "yes" to a number of items would have to be introduced only for the points of concentration. Let us, as an example, assume three points. Let there be ν_{I}, ν_{II}, and ν_{III} proportions of the population concentrated at each point. Designate the latent probabilities pertaining to these three points by p_{I}, p_{II}, and p_{III} respectively. Then the manifest margi-

nals of two items would be related to these latent parameters by the following equations:

$$
\begin{aligned}
p_1 &= \nu_{\mathrm{I}} p_{\mathrm{I}1} + \nu_{\mathrm{II}} p_{\mathrm{II}1} + \nu_{\mathrm{III}} p_{\mathrm{III}1} \\
p_2 &= \nu_{\mathrm{I}} p_{\mathrm{I}2} + \nu_{\mathrm{II}} p_{\mathrm{II}2} + \nu_{\mathrm{III}} p_{\mathrm{III}2} \\
&\nu_{\mathrm{I}} + \nu_{\mathrm{II}} + \nu_{\mathrm{III}} = 1
\end{aligned} \tag{1.9}
$$

Equations (1.9) mean this: The probability of a positive answer for the whole sample is the weighted sum of the segmental probabilities. If we know how many people at each point say "yes" to an item, then we also know how many do so in the whole population.

The trace lines in this simplified model could be looked upon in a variety of ways. In the most rigid sense one might have to say that they, too, consist only of discrete points. More loosely, one might consider them polygons connecting the three points. But for the sake of a more general and consistent algebra, it is preferable to continue to think of them as polynomials. In order to reduce the redundancy of such a model to a minimum the degree of the polynomial will be determined by the number of points. In the case of three classes, for instance, we have for each item three latent probabilities which would determine a polynomial of second degree, a quadratic parabola. In the case of $(m + 1)$ points the degree of the polynomial would be m.

Following exactly the same argument as before, we can then write down the general accounting equations for a "pure test":

$$
p_{ijk\ldots} = \nu_{\mathrm{I}} p_{\mathrm{I}i} p_{\mathrm{I}j} p_{\mathrm{I}k} \cdots + \nu_{\mathrm{II}} p_{\mathrm{II}i} p_{\mathrm{II}j} p_{\mathrm{II}k} \cdots + \nu_{\mathrm{III}} p_{\mathrm{III}i} p_{\mathrm{III}j} p_{\mathrm{III}k} \cdots \tag{1.10}
$$

It is important to realize all the implications of equations (1.10). If we have, let us say, six items, the system would consist of 64 equations. But there are only 18 segmental probabilities and 3 proportions for the segments of the population. Nevertheless, these 64 equations for 21 unknowns do not lead to contradictions. This is due to the postulate that the trace lines are all quadratic parabolas and that, therefore, the manifest data satisfy certain conditions, one of which is that the joint occurrence matrix as discussed above would have the rank 3.

Equations (1.10) have a form quite analogous to equations (1.2). It will be recalled that we multiplied the trace lines $f_i(x)$ and the distribution function $\phi(x)$ and then performed an integration over the x-continuum. Now, we multiply at each point of the x-continuum the segmental probabilities p_x and the population

proportions ν_x and then we perform a finite summation over all points. It is for this reason that these new equations can be called the *equivalent finite accounting equations.*

We have not yet shown the conditions under which a partition exists which has the properties just outlined. This will be done in the next section of this chapter. At this point, we will take its existence for granted and consider what its implication will be for the progress of our discussion. The solution of equations (1.10) would give us the following information:

(1) The coefficients of the polynomial trace lines.

(2) The proportion of the sample which has to be assumed at each point of concentration.

Both these results can be formulated in still another way. Instead of (1) we can say that we would find the latent probabilities corresponding to the concentration points. These probabilities would be the ordinates of m points from which the polynomial trace lines, and therefore their coefficients, would be determined. Instead of (2) we could say that we would know the segmentalization of the population or its discrete distribution. Without going into further detail, it might be added that this whole geometrical picture would be determined but for a linear transformation. The distances between the points of concentration on the x-axis can be drawn arbitrarily.

The segmental probabilities and the proportion of people at each point of the distribution curve we shall call *latent parameters.* The whole system of these latent parameters will be called the *latent structure* of the manifest test material. To the derivation, computation, and use of the latent structure the rest of our two chapters is devoted. We can now conclude our introductory summary as follows:

q. If the polynomial trace lines are of degree m, then a specially manageable model can be developed by assuming that the population is concentrated at $(m + 1)$ points.

r. In this case, finite accounting equations establish a relationship between the manifest data and the underlying continuum on the same logical principle as do the general accounting equations.

s. They do not, however, utilize the probabilities of positive response for each point on the x-continuum; they work with segmental probabilities as would exist for $(m + 1)$ points of the underlying continuum.

t. The ensuing equations are based on finite summations over all points and can be completely solved. (Equations (1.10).)

u. The solutions of these equations are called the latent structure. This latent structure is a determinable equivalent of the original model with which we started. It gives for each test item the trace lines and $(m + 1)$ proportions for the discrete distribution function.

v. The parameters of the model are determined but for a linear transformation on the x-axis in as much as the distances between the points of concentration are arbitrary.

In the present approach two groups of elements can easily be distinguished. On the one side we have *manifest empirical data:* the answers of people to test items, singly and in combinations of items of all order. On the other hand, we present a model whose existence can neither be proved nor disproved directly. We represent the process of test building in terms of *trace lines,* claiming that whatever a test builder considers can be translated into more or less specific assumptions as to trace lines. We also claim that the idea of measuring an underlying continuum with the help of an itemized test can be fully expressed by certain accounting equations. Since such a "formalization" of the process of test building is a postulation not subject to direct verification, it can be judged only in terms of its reasonableness and ultimate usefulness.

Any specific model, however, can be tested. Its parameters will be related by definite equations to the manifest data, on the one hand, and to the general model—the trace lines and the distribution function—on the other hand. These relationships are of a mathematical nature and can be checked by any expert. One such specific case is the latent class structure, which was described in the last paragraph and which will now be explored further in the next sections. In the end we shall come back briefly to the many other types of specifications which could be introduced into the general model.

SECTION II

THE RELATION BETWEEN MANIFEST DATA AND LATENT PARAMETERS

1. THE DICHOTOMOUS SYSTEM

In Section I we have presented the general considerations which lead to the notion of a latent structure and its relation to the analysis of itemized tests. The main idea is simple. A population is

considered divided into a number of latent classes. The answers to test items given by people *within* one latent class are unrelated. But from class to class the proportion of answers to the same item varies. As a consequence, the answers to two or more items *for the whole sample* will appear related. It is a generalization of a well-known fact exemplified by the following number scheme:

$$
\begin{array}{c|c|c|c}
 & + & - & \\
\hline
+ & 12 & 3 & 15 \\
\hline
- & 4 & 1 & 5 \\
\hline
 & 16 & 4 & 20
\end{array}
\quad + \quad
\begin{array}{c|c|c|c}
 & + & - & \\
\hline
+ & 1 & 5 & 6 \\
\hline
- & 3 & 15 & 18 \\
\hline
 & 4 & 20 & 24
\end{array}
\quad = \quad
\begin{array}{c|c|c|c}
 & + & - & \\
\hline
+ & 13 & 8 & 21 \\
\hline
- & 7 & 16 & 23 \\
\hline
 & 20 & 24 & 44
\end{array}
$$

In the first two boxes the answers to two dichotomous items are unrelated: whether people are positive or negative on one item they have the same ratio of positive and negative replies on the second item. But when the two tables are added up box by box, we get the table on the right and here a relationship appears: people who give a positive reply to one item are also more likely to give a positive reply to the other.

Inversely, one can start with the table on the right side and look for ways to find the two tables on the left. For many items and several latent classes this is what latent structure analysis intends to do.

In the present chapter this problem will be treated in a general way. Therefore *one need not think of the latent classes as ordered in any way nor as employing the segments of an underlying continuum.* The more general approach will be connected with test problems in Section VI. There it will be seen that the general solutions to be given now permit us to go beyond the program of the preceding section and to deal with several underlying continua. In the present section stress will be laid not only on solutions achieved but also on work still to be done and especially on aspects of the mathematical problem which are not yet clarified.

Our starting point will always be a set, S, of m dichotomous items such as a series of true-false questions or of attributes an individual might or might not have. As will be shown later, the whole approach can be immediately applied to items which come in more than two classes, for instance, trichotomies.[5]

[5] The letter S here is an abbreviation of the term "system." It should not be confused with the symbol S_i which has appeared in Section I as a computational magnitude formed from manifest data. This magnitude will reappear again in Section III. The discussion of the dichotomous system S is restricted to the present Section II.

The m dichotomies permit 2^m combinations. We could represent the system, S, by these 2^m "response patterns."

But for the present purposes it is better to represent the system, S, by the *joint positive frequencies of all orders*. Let the m items be arranged in an arbitrary order and then give Arabic numbers to them. Let n be the size of the sample, n_1 be the number of people who say "yes" to the first question, n_{12} be the number of people who say "yes" to the first two questions, etc. The number of indexes which such a joint positive frequency has will be called its order. The highest order which can occur is of course m and belongs to the number of people who answer all items positively. The zero-order frequency is identical with the size of the sample.

According to a well-known proposition in combinatorics,[6] the number of these joint frequencies is 2^m. There are, then, as many response patterns as there are joint positive frequencies. The response patterns are all of order m, but include positive as well as negative answers to each item. The positive joint frequencies are of all orders but include only positive replies.

Any theory of scaling aims at giving some kind of order and mutual relationship to the 2^m response patterns of a given set of items. The present theory has been formulated in such a general way that, as far as we can see, it covers all other procedures proposed so far and therefore permits us to establish the relationship between the different scaling systems now extant. While we shall return to response patterns later, it will be necessary first to discuss at some length the mathematical properties of the joint positive frequencies.

2. THE REDUCTION OF A DICHOTOMOUS SYSTEM

The relative positive frequencies for a series of dichotomies will be indicated by the letter p with an Arabic numeral subscript indicating to which dichotomous item it refers; relative negative frequencies will likewise be indicated by the letter q. We have therefore the following symbolic conventions:

$$\frac{n_1}{n} = p_1, \qquad \frac{n_2}{n} = p_2, \text{ etc.}$$

$$q_1 = 1 - p_1, \qquad q_2 = 1 - p_2, \text{ etc.}$$

According to the usage of empirical social research, these will be called relative marginals.

[6] See G. U. Yule and M. J. Kendall, *An Introduction to the Theory of Statistics* (12th ed., J. B. Lippincott Co., Philadelphia, 1940), page 17.

Definition: A homogeneous dichotomous system is one where the higher order frequencies can be computed if the size of the sample and the relative marginals (the p_i's and the q_i's) are known. The procedure of computation is evident from the following formulae:

$$\frac{n_{12}}{n} = p_1p_2, \qquad \frac{n_{123}}{n} = p_1p_2p_3, \qquad \frac{n_{23}}{n} = p_2p_3, \text{ etc.}$$

This means nothing else, of course, but that in a homogeneous S the items are independent of each other. In terms of probabilities, the higher order probabilities are the product of the first order probabilities.

In empirical research, when a system of items is unrelated in this way there is, so to speak, nothing further to explain. If items are relatéd, then explanation usually consists of introducing additional items so that for certain subsamples the relationships disappear. This well-known procedure of successive cross tabulations is formalized here by the following definition:

Definition: A dichotomous S of m items is called reducible into λ strata if the following conditions prevail:

(a) There exist λ homogeneous systems of m items.

(b) A unique correspondence is established between each frequency in the original system and one frequency in each of the homogeneous systems.

(c) By adding corresponding items in the homogeneous systems we obtain the corresponding item in the original system.

The relationship between a reducible system, S, and its component strata will be symbolized by the following equation:

$$S = S_{\text{I}} + S_{\text{II}} + \cdots + S_{\iota}$$

To illustrate this definition we present in Table 1 a simple example where $m = 4$ and $\lambda = 2$. The systems are given in two forms: (A) through their sixteen response patterns and (B) through their sixteen joint positive frequencies. The arrangement of the scheme makes clear the kind of correspondence which exists between the various systems. It is worth remembering that the frequency n_{1234} is equal to that for response pattern $+ + + +$, and for example $n_{234} - n_{1234}$ gives the frequency of response pattern $- + + +$.[7]

[7] See Yule and Kendall, *op.cit.*, page 19, for rules of computation.

The frequencies in the homogeneous systems we call "latent" and the original frequencies "manifest." The reader should satisfy himself on two counts. The sum of corresponding latent frequencies is the corresponding manifest frequency. And the two latent systems are homogeneous; that is, the higher order frequencies are the products of the first order frequencies. From part B of Table 1, the reader may check the relation between n_{12} and $n_1 n_2$. In the second and third columns (in the two latent subclasses) n_{12} =

TABLE 1

A Reduction for the Case $m = 4$, $\lambda = 2$

1	*2*	*3*	*4*	*Total*	*Latent class I*	*Latent class II*
				A. IN TERMS OF RESPONSE PATTERNS		
+	+	+	+	72.2	71.9	.3
+	+	+	−	72.8	68.4	4.4
+	+	−	+	56.8	54.2	2.6
+	−	+	+	44.0	42.7	1.3
−	+	+	+	6.3	5.9	.4
+	+	−	−	90.8	51.6	39.2
+	−	+	−	59.7	40.6	19.1
+	−	−	+	43.7	32.2	11.5
−	+	+	−	11.4	5.6	5.8
−	+	−	+	8.0	4.5	3.5
−	−	+	+	5.2	3.5	1.7
+	−	−	−	201.0	30.6	170.4
−	+	−	−	55.7	4.2	51.5
−	−	+	−	28.4	3.3	25.1
−	−	−	+	17.8	2.6	15.2
−	−	−	−	226.2	2.5	223.7
				B. IN TERMS OF JOINT FREQUENCIES		
n				1,000	424.3	575.7
n_1				641	392.2	248.8
n_2				374	266.3	107.7
n_3				300	241.9	58.1
n_4				254	217 5	36.5
n_{12}				292.6	246.1	46.5
n_{13}				248.7	223.6	25.1
n_{14}				216.7	201.0	15.7
n_{23}				162.7	151.8	10.9
n_{24}				143.3	136.5	6.8
n_{34}				127.7	124.0	3.7
n_{123}				145.0	140.3	4.7
n_{124}				129.0	126.1	2.9
n_{134}				116.2	114.6	1.6
n_{234}				78.5	77.8	.7
n_{1234}				72.2	71.9	.3

$\frac{n_1 n_2}{n}$. But in the first column (representing the manifest data for the whole sample) this is not the case. Still, row by row, the first column is the sum of the other two.

A remark is in order on the similarity between this procedure of reduction and factor analysis. Habitually we say that test scores are the linear sum of the factors. But we can also say that the correlation between scores is zero when the factors are partialled out. The latent systems in our present scheme can be looked upon as relationships between items which have a "zero correlation" if the latent structure is partialled out. We shall come back to this interpretation in considerable detail. At the moment, however, we still remain in the realm of algebra, and raise the following question: Under what conditions can a given system, S, of m items be reduced to λ homogeneous subsystems?

3. THE LATENT STRUCTURE

Let us assume that a reduction has successfully been carried out for a system of $\lambda = 3$ and $m = 6$. Then we would have the following information:

(a) We would know the number of cases in each of the three subclasses: the *latent class frequencies.*

(b) We would know for each subclass the six "*latent marginals,*" that is, the proportion of people who give a positive answer to each of the six items.

(c) From these latent marginals we could get all the higher order frequencies in each subclass.

(d) By adding up corresponding latent frequencies we could get the corresponding manifest frequencies.

The *latent structure* is summarized in Table 2. The *latent structure equations* shown in Table 3 are derived from this scheme by multiplications within each row and additions over the rows in Table 2. On the left of the latent structure equations are the joint frequencies as they are observed from an actual test. On the right are latent parameters which are unknown to begin with and which we desire to compute with the help of the equations.

In Table 2 the latent subsystems are put into an arbitrary order and indicated by Roman numerals. In the scheme each row corresponds to a subsystem and each column to an item. The relation between latent and manifest frequencies is given by $2^6 = 64$ equa-

TABLE 2

LATENT STRUCTURE FOR $m = 6$ AND $\lambda = 3$

Latent class frequencies	*Latent marginals for items* 1	2	3	4	5	6
n_I	p_{I1}	p_{I2}	p_{I3}	p_{I4}	p_{I5}	p_{I6}
n_{II}	p_{II1}	p_{II2}	p_{II3}	p_{II4}	p_{II5}	p_{II6}
n_{III}	p_{III1}	p_{III2}	p_{III3}	p_{III4}	p_{III5}	p_{III6}

TABLE 3

THE LATENT STRUCTURE EQUATIONS

$$n = n_I + n_{II} + n_{III} = \sum n_\nu$$

$$n_1 = n_I p_{I1} + n_{II} p_{II1} + n_{III} p_{III1} = \sum n_\nu p_{\nu 1}$$

$$n_2 = n_I p_{I2} + n_{II} p_{II2} + n_{III} p_{III2} = \sum n_\nu p_{\nu 2}$$

$$\vdots$$

$$n_{12} = n_I p_{I1} p_{I2} + n_{II} p_{II1} p_{II2} + n_{III} p_{III1} p_{III2} = \sum n_\nu p_{\nu 1} p_{\nu 2}$$

$$n_{13} = n_I p_{I1} p_{I3} + n_{II} p_{II1} p_{II3} + n_{III} p_{III1} p_{III3} = \sum n_\nu p_{\nu 1} p_{\nu 3}$$

$$\vdots$$

$$n_{123} = n_I p_{I1} p_{I2} p_{I3} + n_{II} p_{II1} p_{II2} p_{II3} + n_{III} p_{III1} p_{III2} p_{III3} = \sum n_\nu p_{\nu 1} p_{\nu 2} p_{\nu 3}$$

$$n_{124} = n_I p_{I1} p_{I2} p_{I4} + n_{II} p_{II1} p_{II2} p_{II4} + n_{III} p_{III1} p_{III2} p_{III4} = \sum n_\nu p_{\nu 1} p_{\nu 2} p_{\nu 4}$$

$$\vdots$$

$$n_{123456} = \cdots\cdots\cdots = \sum n_\nu p_{\nu 1} p_{\nu 2} p_{\nu 3} p_{\nu 4} p_{\nu 5} p_{\nu 6}$$

tions. The reader is invited to familiarize himself with the scheme of Table 2 and 3, since it will be the basis for all further analysis.

In general, the number of parameters in a latent structure is $\lambda(m + 1)$. In each of the λ subsystems we have m latent marginals and to this we have to add the total number of cases or latent class frequencies, one for each subsystem.

These $\lambda(m + 1)$ latent parameters are linked by 2^m equations. A reduction, therefore, will in general not be possible if $2^m > \lambda(m + 1)$ except if certain additional conditions are satisfied by the manifest frequencies so that the number of independent equations

is reduced to at least the number of latent parameters. In order to derive these conditions we have to discuss once more the presentation of the system S itself.

4. THE JOINT FREQUENCY MATRICES

In Table 4 we have a symmetric matrix where the left upper corner stands for the size of the sample. The first row and the first column are filled by the manifest marginals of the items of the system. In the boxes outside the main diagonal we have the number of people who give positive answers to any pair of two items which can be formed. In the diagonal boxes we have symbols in pairs which so far have no meaning and are not defined in the system, S. They will hereafter be called the "missing frequencies." We shall comment upon them presently.

TABLE 4

THE MANIFEST JOINT OCCURRENCE MATRIX M

		1	*2*	*3*	*4*	*5*	*6*
	n	n_1	n_2	n_3	n_4	n_5	n_6
1	n_1	(n_{11})	n_{12}	n_{13}	n_{14}	n_{15}	n_{16}
2	n_2	n_{12}	(n_{22})	n_{23}	n_{24}	n_{25}	n_{26}
3	n_3	n_{13}	n_{23}	(n_{33})	n_{34}	n_{35}	n_{36}
4	n_4	n_{14}	n_{24}	n_{34}	(n_{44})	n_{45}	n_{46}
5	n_5	n_{15}	n_{25}	n_{35}	n_{45}	(n_{55})	n_{56}
6	n_6	n_{16}	n_{26}	n_{36}	n_{46}	n_{56}	(n_{66})

Let us suppose now that we have stratified the sample by item 1 and picked out all the people who give a positive answer to item 1. We then get the frequency with which these people give a positive reply to all possible pairs of remaining items. These frequencies properly inserted give M^1. In this matrix the row and column of item 1 as well as the main diagonal will consist of missing frequencies. We shall call this a joint occurrence matrix of signature 1. Correspondingly, we shall have such a matrix for any other "signature." M^{23}, e.g., will originate in the following way: we first pick all the people who make positive responses to items 2 and 3; then we find by cross tabulations the number of these people who jointly make a positive response to the pairs of remaining items. This matrix has signature 23. These joint occurrences of higher order are then arranged in a way exemplified by Table 5.

The system S can be presented by the totality of such matrices. This presentation is redundant. Some of the manifest frequencies in the first line of M^{23}, e.g., also appear in M^{34}, in M^{24}, in M^{35}, etc. If the M-matrices are arranged in lexical order, the rule is this: The rows and columns corresponding to the signature of M consist of missing frequencies only. All "preceding" rows and columns have already appeared in a "previous" M of the same order but of an "earlier" signature.

TABLE 5

THE JOINT OCCURRENCE MATRIX M^{23}

		1	*2*	*3*	*4*	*5*	*6*
	n_{23}	n_{123}	(n_{223})	(n_{233})	n_{234}	n_{235}	n_{236}
1	n_{123}	(n_{1123})	(n_{1223})	(n_{1233})	n_{1234}	n_{1235}	n_{1236}
2	(n_{223})	(n_{1223})	(n_{2223})	(n_{2233})	(n_{2234})	(n_{2235})	(n_{2236})
3	(n_{233})	(n_{1233})	(n_{2233})	(n_{2333})	(n_{2334})	(n_{2335})	(n_{2336})
4	n_{234}	n_{1234}	(n_{2234})	(n_{2334})	(n_{2344})	n_{2345}	n_{2346}
5	n_{235}	n_{1235}	(n_{2235})	(n_{2335})	n_{2345}	(n_{2355})	n_{2356}
6	n_{236}	n_{1236}	(n_{2236})	(n_{2336})	n_{2346}	n_{2356}	(n_{2366})

In spite of this redundance, the joint occurrence matrices will turn out to be a powerful tool of analysis when we want to study the reduction of S in more detail.

5. THE RELATION BETWEEN THE LATENT STRUCTURE AND THE MATRICES OF MANIFEST JOINT FREQUENCIES

Let (N) be a diagonal matrix the elements of which are the latent class frequencies. For $\lambda = 3$, for instance, we have

$$(N) = \begin{Vmatrix} n_{\mathrm{I}} & 0 & 0 \\ 0 & n_{\mathrm{II}} & 0 \\ 0 & 0 & n_{\mathrm{III}} \end{Vmatrix}$$

Let L be a matrix the elements of which are the latent marginals bordered by a column of unities. For $m = 6$, $\lambda = 3$, we have

$$L = \begin{Vmatrix} 1 & p_{\mathrm{I}1} & p_{\mathrm{I}2} & p_{\mathrm{I}3} & p_{\mathrm{I}4} & p_{\mathrm{I}5} & p_{\mathrm{I}6} \\ 1 & p_{\mathrm{II}1} & p_{\mathrm{II}2} & p_{\mathrm{II}3} & p_{\mathrm{II}4} & p_{\mathrm{II}5} & p_{\mathrm{II}6} \\ 1 & p_{\mathrm{III}1} & p_{\mathrm{III}2} & p_{\mathrm{III}3} & p_{\mathrm{III}4} & p_{\mathrm{III}5} & p_{\mathrm{III}6} \end{Vmatrix}$$

The matrix L has λ rows and $(m + 1)$ columns in the general case. It is practically a replica of the latent structure scheme.

Finally, we shall, by the usual convention, designate by L' the transposed matrix L. Then some of the equations of Table 3 can be put in the following form:

$$M = L'(N)\,L$$

(The reader in checking this formula will remember that premultiplying a matrix by a diagonal matrix means to multiply each column of the matrix by the element in the corresponding column of the diagonal matrix.)

New *N-matrices of higher signature* will be introduced by the following types of defining equations:

$$(N^1) = \begin{Vmatrix} n_{\mathrm{I}}p_{\mathrm{I}1} & 0 & 0 \\ 0 & n_{\mathrm{II}}p_{\mathrm{II}1} & 0 \\ 0 & 0 & n_{\mathrm{III}}p_{\mathrm{III}1} \end{Vmatrix}$$

$$(N^{12}) = \begin{Vmatrix} n_{\mathrm{I}}p_{\mathrm{I}1}p_{\mathrm{I}2} & 0 & 0 \\ 0 & n_{\mathrm{II}}p_{\mathrm{II}1}p_{\mathrm{II}2} & 0 \\ 0 & 0 & n_{\mathrm{III}}p_{\mathrm{III}1}p_{\mathrm{III}2} \end{Vmatrix} \quad \text{etc.}$$

They are all diagonal matrices. Their diagonal elements each correspond successively to one of the latent subsystems. Their "signature" indicates how many and which items are involved in the formation of the matrix. Each element is the product of a latent class frequency and an indicated number of latent marginals. The element in the third row of (N^{235}), e.g., will be $n_{\mathrm{III}}p_{\mathrm{III}2}p_{\mathrm{III}3}p_{\mathrm{III}5}$.

It then can easily be seen that equations of the following type hold true:

$$M^1 = L'(N^1)L$$
$$M^{12} = L'(N^{12})L$$
$$\text{etc.}$$

All the equations in Table 3, which link the latent structure and the manifest positive frequencies in the system S, can therefore be summarized in the *basic matrix equation of latent structure analysis:*

$$M^\sigma = L'(N^\sigma)L \qquad \sigma = 0, 1, 2, 3, \cdots, i, 12, 13, \cdots$$

where σ is the signature of the manifest M-matrix and the latent N-matrices, and $M^0 = M$, of course.

6. THE BASIC DETERMINANTS AND THEIR PROPERTIES

From the basic matrix equation we can derive a variety of more specific equations by forming what will be called *basic determinants* of the M-matrices. We shall call a *basic determinant* of the matrix M any determinant which includes the element n and some other elements of the first row and of the first column, but no element of the main diagonal (other than the element n). Such a determinant is formed by picking certain additional rows and columns. Once they are indicated, the whole determinant is known. Therefore, the following symbolism is self-explanatory. The first group of Arabic figures indicates the additional columns selected, and the second group the additional rows. Thus:

$$\begin{vmatrix} n & n_1 \\ n_2 & n_{12} \end{vmatrix} = [1,2] \quad , \quad \begin{vmatrix} n & n_2 & n_5 \\ n_1 & n_{12} & n_{15} \\ n_3 & n_{23} & n_{35} \end{vmatrix} = [2\ 5, 1\ 3]$$

If basic determinants are formed from an M-matrix with higher signature than zero, this will be indicated after a semicolon. From M^2, for instance we could form

$$[1\ 4, 5\ 6; 2] = \begin{vmatrix} n_2 & n_{21} & n_{24} \\ n_{25} & n_{215} & n_{245} \\ n_{26} & n_{216} & n_{246} \end{vmatrix}$$

The basic matrix equations permit us to express these basic determinants in terms of latent parameters. These expressions have properties which can either be proved directly or derived from a general theorem.[8]

For our purpose, it is best to single out five properties. They will be formulated in general terms and exemplified for the case $\lambda = 3$.

(a) *Basic determinants of order higher than λ will always vanish.* For instance,

---

[8] A. C. Aitken, *Determinants and Matrices* (Oliver and Boyd, Ltd., Edinburgh, 1944), page 85.

$$
(2.1) \qquad [1\ 2\ 3,\ 4\ 5\ 6] = \begin{vmatrix} n & n_1 & n_2 & n_3 \\ n_4 & n_{14} & n_{24} & n_{34} \\ n_5 & n_{15} & n_{25} & n_{35} \\ n_6 & n_{16} & n_{26} & n_{36} \end{vmatrix}
$$

$$
= \begin{vmatrix} \sum n_i & \sum n_i p_{i1} & \sum n_i p_{i2} & \sum n_i p_{i3} \\ \sum n_i p_{i4} & \sum n_i p_{i1} p_{i4} & \sum n_i p_{i2} p_{i4} & \sum n_i p_{i3} p_{i4} \\ \sum n_i p_{i5} & \sum n_i p_{i1} p_{i5} & \sum n_i p_{i2} p_{i5} & \sum n_i p_{i3} p_{i5} \\ \sum n_i p_{i6} & \sum n_i p_{i1} p_{i6} & \sum n_i p_{i2} p_{i6} & \sum n_i p_{i3} p_{i6} \end{vmatrix}
$$

$$
= n_{\mathrm{I}} n_{\mathrm{II}} n_{\mathrm{III}} \begin{vmatrix} 1 & 1 & 1 & 0 \\ p_{\mathrm{I}4} & p_{\mathrm{II}4} & p_{\mathrm{III}4} & 0 \\ p_{\mathrm{I}5} & p_{\mathrm{II}5} & p_{\mathrm{III}5} & 0 \\ p_{\mathrm{I}6} & p_{\mathrm{II}6} & p_{\mathrm{III}6} & 0 \end{vmatrix} \cdot \begin{vmatrix} 1 & p_{\mathrm{I}1} & p_{\mathrm{I}2} & p_{\mathrm{I}3} \\ 1 & p_{\mathrm{II}1} & p_{\mathrm{II}2} & p_{\mathrm{II}3} \\ 1 & p_{\mathrm{III}1} & p_{\mathrm{III}2} & p_{\mathrm{III}3} \\ 0 & 0 & 0 & 0 \end{vmatrix} = 0
$$

The first part of (2.1) may be derived from the scheme of Table 3; the second part is a simple application of matrix multiplication.

(b) *Basic determinants of order λ can be presented as the product of three other determinants of order λ, each one taken respectively and correspondingly from one of the three matrices L', N, L.* For instance,

$$
[1\ 2,\ 3\ 4] = \begin{vmatrix} n & n_1 & n_2 \\ n_3 & n_{13} & n_{23} \\ n_4 & n_{14} & n_{24} \end{vmatrix} = \begin{vmatrix} \sum n_i & \sum n_i p_{i1} & \sum n_i p_{i2} \\ \sum n_i p_{i3} & \sum n_i p_{i1} p_{i3} & \sum n_i p_{i2} p_{i3} \\ \sum n_i p_{i4} & \sum n_i p_{i1} p_{i4} & \sum n_i p_{i2} p_{i4} \end{vmatrix}
$$

$$
(2.2) \qquad = \begin{vmatrix} 1 & 1 & 1 \\ p_{\mathrm{I}3} & p_{\mathrm{II}3} & p_{\mathrm{III}3} \\ p_{\mathrm{I}4} & p_{\mathrm{II}4} & p_{\mathrm{III}4} \end{vmatrix} n_{\mathrm{I}} n_{\mathrm{II}} n_{\mathrm{III}} \begin{vmatrix} 1 & p_{\mathrm{I}1} & p_{\mathrm{I}2} \\ 1 & p_{\mathrm{II}1} & p_{\mathrm{II}2} \\ 1 & p_{\mathrm{III}1} & p_{\mathrm{III}2} \end{vmatrix}
$$

If the M-matrix is of higher signature, this will show up in the product which corresponds to the N-matrix. For instance,

$$
[1\ 2,\ 3\ 4;\ 5] = \begin{vmatrix} 1 & 1 & 1 \\ p_{\mathrm{I}3} & p_{\mathrm{II}3} & p_{\mathrm{III}3} \\ p_{\mathrm{I}4} & p_{\mathrm{II}4} & p_{\mathrm{III}4} \end{vmatrix} n_{\mathrm{I}} p_{\mathrm{I}5} n_{\mathrm{II}} p_{\mathrm{II}5} n_{\mathrm{III}} p_{\mathrm{III}5} \begin{vmatrix} 1 & p_{\mathrm{I}1} & p_{\mathrm{I}2} \\ 1 & p_{\mathrm{II}1} & p_{\mathrm{II}2} \\ 1 & p_{\mathrm{III}1} & p_{\mathrm{III}2} \end{vmatrix}
$$

It is important to realize that this will be true only for basic determinants of order λ, when λ is the number of latent subclasses.

(c) *The ratio of two basic determinants of order λ taken correspondingly from a matrix of higher signature M^{σ} and the matrix M*

depends only upon the items kept constant for the formation of M^{σ} *and not upon the items involved in the formation of the determinants.*

From the preceding example we see immediately that

$$\frac{[1\,2,\,3\,4;\,5]}{[1\,2,\,3\,4]} = p_{\mathrm{I}5}p_{\mathrm{II}5}p_{\mathrm{III}5} \tag{2.3a}$$

Because of the importance of this ratio for practical computations, we shall give it a special symbol μ_i, where the index refers to the item whose latent marginals are $p_{\mathrm{I}i}p_{\mathrm{II}i}p_{\mathrm{III}i}$; in this case we have, for instance, $\mu_5 = p_{\mathrm{I}5}p_{\mathrm{II}5}p_{\mathrm{III}5}$.

In the same way we would find, e.g., that

$$\frac{[a\,b,\,c\,d;\,2\,3]}{[a\,b,\,c\,d]} = p_{\mathrm{I}2}p_{\mathrm{II}2}p_{\mathrm{III}2}p_{\mathrm{I}3}p_{\mathrm{II}3}p_{\mathrm{III}3} = \mu_{23} \tag{2.3b}$$

where a, b, c, and d indicate any items other than 2 or 3.

The ratio between such corresponding determinants of order λ will be called their *link*. It has great importance for practical computations as well as for the whole theory.

(d) *Links are multiplicative* in the sense that

$$\mu_{ijk}\ldots. = \mu_i\mu_j\mu_k\cdots \tag{2.4a}$$

This can be seen immediately from the formula just developed. The result can be expressed in still another way. We can see that, e.g.,

$$\frac{\mu_{ijk}}{\mu_{ij}} = \mu_k \tag{2.4b}$$

This means that the result (c) can be generalized in the following way: The ratio of two basic determinants of order λ taken correspondingly from two M-matrices of different signature depends only upon the signature of the two matrices and not upon the items involved in the formation of the determinants.

It will help to clarify the meaning of (2.3) and (2.4) if they are exemplified by the simple case $\lambda = 2$, and $m = 5$. According to (2.3) we have, e.g.

$$\frac{[12;5]}{[12]} = \frac{[34;5]}{[34]}$$

This means we stratify a sample according to its answer to item 5. Then the "relationships" between items 1 and 2 might be lower than

they were for the whole sample. This "drop" is expressed by the ratio $\mu_5 = \frac{[12;5]}{[12]}$. Keeping item 5 constant also introduces a "drop" in all other cross products, e.g., $\frac{[34;5]}{[34]}$. What item 5 "takes out" of the relationships between any two other items depends upon the "stratifier" (item 5) only and is the same for all ratios $\frac{[ij;5]}{[ij]}$.

(2.4) states that $\frac{[12;45]}{[12]} = \frac{[12;4]}{[12]} \times \frac{[12;5]}{[12]}$. The "drop" induced into a relationship [12] by two simultaneous stratifications (according to items 4 and 5) is the product of the "drops" induced by the two stratifications induced independently from each other. In its alternative form, (2.4) indicates that:

$$\frac{[12;45]}{[12;4]} = \frac{[12;5]}{[12]}$$

This means that stratification by item 5 induces a drop in the relationship between items 1 and 2 which is the same, irrespective of the level of stratification performed previously with the help of other items. In a somewhat more general formulation:

$$\frac{[ij;kl5]}{[ij;kl]} = \frac{[ij;k5]}{[ij;k]} = \frac{[ij;5]}{[ij]}$$

(e) *Determinants of orders less than* λ take a somewhat more complicated form. Just for the record the following example is added for the case $\lambda = 3$.

$$(2.5a)\quad [1, 2] = \begin{vmatrix} n & n_1 \\ n_2 & n_{12} \end{vmatrix} = n_{\mathrm{I}}n_{\mathrm{II}} \begin{vmatrix} 1 & 1 \\ p_{\mathrm{I}2} & p_{\mathrm{II}2} \end{vmatrix} \cdot \begin{vmatrix} 1 & p_{\mathrm{I}1} \\ 1 & p_{\mathrm{II}1} \end{vmatrix}$$
$$+ n_{\mathrm{I}}n_{\mathrm{III}} \begin{vmatrix} 1 & 1 \\ p_{\mathrm{I}2} & p_{\mathrm{III}2} \end{vmatrix} \cdot \begin{vmatrix} 1 & p_{\mathrm{I}1} \\ 1 & p_{\mathrm{III}1} \end{vmatrix} + n_{\mathrm{II}}n_{\mathrm{III}} \begin{vmatrix} 1 & 1 \\ p_{\mathrm{II}2} & p_{\mathrm{III}2} \end{vmatrix} \cdot \begin{vmatrix} 1 & p_{\mathrm{II}1} \\ 1 & p_{\mathrm{III}1} \end{vmatrix}$$

If $\lambda = 2$ this reduces to

$$(2.5b)\qquad [ij] = n_{\mathrm{I}}n_{\mathrm{II}}\,(p_{\mathrm{I}i} - p_{\mathrm{II}i})\,(p_{\mathrm{I}j} - p_{\mathrm{II}j})$$

The formulae (2.5) which give the ordinary cross product in terms of the latent parameters will turn out to be especially useful for computational purposes.

7. CONDITIONS FOR REDUCIBILITY

It is now simple to formulate the properties which a system, S, has to have in order to be reducible to a "sum" of λ homogeneous systems.

(a) *The rank of all joint occurrence matrices has to be* λ. From proposition (a) in the previous paragraph, it is clear that the rank cannot be larger than λ; in a variety of ways, preferably from the definition of "links," it can be seen that the rank cannot be smaller than λ. The product of the latent marginals would become indefinite if all basic determinants of order λ were zero. A similar conclusion can be drawn from the actual computation of the latent parameters, to be discussed presently.

(b) The ratio between corresponding basic determinants of order λ taken from two matrices M^{σ} and M should have the same value, whatever items within M^{σ} are used to form these ratios. Or, in other words, *"links" should depend only upon the signature of a stratified joint occurrence matrix.*

(c) *Links should be multiplicative* in the sense indicated in the previous paragraph under finding (d).

There are then essentially three types of conditions which a reducible system has to fulfill. One type (a) takes place *within* all M-matrices. The second type (b) takes place *between* any matrix of higher signature and M, the matrix of joint paired occurrences. The third type (c) takes place between M and two or more matrices of higher signature. Not all of these conditions are independent. It would probably be desirable to show generally that the number of independent conditions is $2^m - (m + 1)\lambda$. This is a rather laborious problem of combinatorics. It has been carried through for the specific cases of $\lambda = 2$ and $\lambda = 3$, but not yet in the general case.

For many purposes it is useful to give the conditions of reducibility in a somewhat different form, by introducing *cross-product matrices*. Take any basic determinant formed from the matrix M for instance:

$$\begin{vmatrix} n & n_1 & n_2 & n_3 \\ n_4 & n_{14} & n_{24} & n_{34} \\ n_5 & n_{15} & n_{25} & n_{35} \\ n_6 & n_{16} & n_{26} & n_{36} \end{vmatrix} = [1\,2\,3,\,4\,5\,6] \qquad (2.6a)$$

By pivotal condensation around the left upper element we can give this determinant a new form:

$$(2.6b) \qquad \frac{1}{n}\begin{vmatrix} [1,4] & [2,4] & [3,4] \\ [1,5] & [2,5] & [3,5] \\ [1,6] & [2,6] & [3,6] \end{vmatrix}$$

where the symbols $[i,j]$ are the well-known cross products of all the fourfold tables which can be formed between the positive and negative occurrences of any pair of items.

Now this form has an analogy to a correlation matrix. Its rank is obviously one less than the original form. The tetrads in this new form are third order determinants in the original form. For instance,

$$(2.6c) \qquad [1,4]\ [2,5] - [1,5]\ [2,4] = n \begin{vmatrix} n & n_1 & n_2 \\ n_4 & n_{14} & n_{24} \\ n_5 & n_{15} & n_{25} \end{vmatrix} = n\,[1\,2, 4\,5]$$

These tetrads will vanish in the case where $\lambda = 2$. In other words, instead of using the original matrix M we could use its condensed form, the cross-product matrix. Its rank would be $r = \lambda - 1$ in a system which is reducible to a "sum" of λ homogeneous systems.

At this point let us return to the "missing frequencies." Symbols like n_{ii} were originally not defined. But from the general matrix formulation at the end of paragraph 5 in this section, a definition enforces itself. In the case of three latent subclasses, for instance, it would have to be as follows:

$$n_{ii} = n_{\text{I}}p_{\text{I}i}^2 + n_{\text{II}}p_{\text{II}i}^2 + n_{\text{III}}p_{\text{III}i}^2$$

This equation is not an empty formalism. Once the latent parameters on its right side are computed from manifest data, we can then compute the numerical values of the missing frequencies. An intuitive interpretation comes close to the notion of reliability: the number of people who in two successive interviews would answer the same question twice in a positive way; this assumes, of course, that the whole test is truly "pure" and the repeated answers are determined only by the class position of the respondents and by otherwise random elements. It might be added that from a further analysis of these missing frequencies are derived additional conditions of reducibility of a rather complicated nature. Because they will not be needed in our subsequent discussion, they will not be considered further here.

It is important to see the conditions of reducibility against the background of the whole theory. We started out in Section I with a model, which consisted of a trace line for each item and a distribution function for the whole population. We found that from the type of manifest data which come from an itemized test, the coefficients of the model cannot be mathematically deduced without further assumptions. Among the various specifications which are conceivable we picked the one whose algebra was developed in the preceding pages: the latent structure which was equivalent to a concentration of the population at λ points and polynomial trace lines of degree $(\lambda - 1)$.

The problem, therefore, shifted to the computation of the latent structure from manifest data. To solve this task, we proceeded as follows: We postulated latent structures with general parameters, and inquired as to the manifest data which such structures would generate. It turned out that a given latent structure imposes certain conditions on the manifest data which could correspond to it. These restrictions were analyzed in the preceding paragraph.

Then we reversed the process. We asked how we could recognize, from an inspection of manifest data, which corresponding latent structure might, as it were, have generated it. Thus we came to develop the present conditions of reducibility. But one point has still to be brought out. A latent structure can lead to only one unique set of manifest data. But the reverse is not true. A given set of data can at least theoretically come from different latent structures. This is due to the possibility of "degenerate" latent structures. We can easily imagine a structure of λ subclasses, where the latent marginals are subject to certain conditions. Such restrictions on the latent structure will express themselves as additional conditions imposed upon the manifest data: additional in the sense of going beyond the general ones developed in the preceding paragraph. The effect will usually be that certain additional determinants in the manifest joint occurrence matrices vanish. In certain cases, these latent restrictions themselves can be computed from the manifest data. In others, this is not possible. A latent structure with λ classes without restriction on the marginals might generate the same manifest data as does a structure with more than λ classes but with conditions imposed upon the latent marginals. This leads to the distinction of "partially degenerate" and "fully degenerate" structures. We shall give here one example of the second possibility. The first possibility, that the latent restrictions

can themselves be discovered from the manifest data, will be exemplified presently.

Take three latent subclasses. Suppose that all items have the same latent marginals in two subclasses, e.g.

$$p_{\mathrm{I}i} = p_{\mathrm{II}i}$$

Then it can easily be seen that all the conditions are satisfied which would lead one to assume that we are dealing with a latent dichotomy. All the tetrads, for instance, will vanish. This can be seen if one computes a cross product by using formula (2.5a) in the form

$$(2.7)\ [ij] = n_{\mathrm{I}}n_{\mathrm{II}}\,(p_{\mathrm{I}i} - p_{\mathrm{II}i})\,(p_{\mathrm{I}j} - p_{\mathrm{II}j}) + n_{\mathrm{I}}n_{\mathrm{III}}\,(p_{\mathrm{I}i} - p_{\mathrm{III}i})(p_{\mathrm{I}j} - p_{\mathrm{III}j}) + n_{\mathrm{II}}n_{\mathrm{III}}\,(p_{\mathrm{II}i} - p_{\mathrm{III}i})\,(p_{\mathrm{II}j} - p_{\mathrm{III}j})$$

By assumption the latent marginals in class I and II are alike for all items. Therefore on the right of the last equation the first element vanishes and the sum of the second two becomes:

$$[ij] = (p_{\mathrm{I}i} - p_{\mathrm{III}i})\,(p_{\mathrm{I}j} - p_{\mathrm{III}j})\,n_{\mathrm{III}}\,(n_{\mathrm{I}} + n_{\mathrm{II}})$$

As can easily be seen therefore any tetrad formed from four items would vanish:

$$\begin{vmatrix} [ij] & [ik] \\ [lj] & [lk] \end{vmatrix} = 0$$

In a similar way it can be shown that all other conditions of reducibility to two latent classes would be satisfied. The existence of three subclasses could be detected only by adding further items which do not have the restrictive property. Inversely, therefore: If the conditions of reducibility to λ subclasses are fulfilled, the latent structure is either a general one for λ subclasses or a fully degenerate one for more than λ subclasses. Which of the two assumptions should be made cannot be decided on mathematical grounds alone. However, if the latter assumption is made, then it is not possible to compute all the latent parameters which a latent structure with more than λ subclasses would require: further empirical material would be needed for this computation. It is obvious how these findings could be generalized to cases with $\lambda > 3$. Not much would be gained if one were to develop the notation needed in order to cover all possible "degenerate" cases. Some of them will come up presently.

SECTION III

THE DERIVATION OF SPECIFIC LATENT STRUCTURES

1. THE GENERAL PROBLEM

The solution of the latent structure equations can be a laborious task, which becomes more difficult as the number of subclasses increases. No general solution has yet been found, and numerical procedures have been developed for special cases only. The method of attack may be exemplified by the case $\lambda = 3$ and $m = 4$. In this chapter it will be assumed that the conditions of reducibility are perfectly fulfilled. In the next chapter we shall deal with formulae for handling actual data, where these conditions are only approximately fulfilled.

Consider, illustratively, the latent structure of Table 6, where the four test items are indicated by Arabic numbers and the three latent subclasses by Roman numbers.

TABLE 6

GENERAL LATENT STRUCTURE FOR $m = 4$, $\lambda = 3$

Latent class frequencies	*Latent marginals for items* 1	2	3	4
n_{I}	$a_1 + c_1$	$a_2 + c_2$	$a_3 + c_3$	$a_4 + c_4$
n_{II}	$a_1 + b_1$	$a_2 + b_2$	$a_3 + b_3$	$a_4 + b_4$
n_{III}	a_1	a_2	a_3	a_4

The latent marginals are now given in a different notation, which for practical applications has two advantages. It dispenses with double indices and puts differences of latent marginals in a simpler form:

$$p_{\text{I}i} - p_{\text{III}i} = c_i$$
$$p_{\text{II}i} - p_{\text{III}i} = b_i$$
$$p_{\text{I}i} - p_{\text{II}i} = c_i - b_i$$

These differences become important because our departure will be formula (2.7). The cross product between two items then becomes:

$$(3.1) \quad [ij] = n_{\text{I}}n_{\text{III}}c_ic_j + n_{\text{II}}n_{\text{III}}b_ib_j + n_{\text{I}}n_{\text{II}}(c_i - b_i)(c_j - b_j) =$$
$$c_ic_jn_{\text{I}}(n - n_{\text{I}}) + b_ib_jn_{\text{II}}(n - n_{\text{II}}) - n_{\text{I}}n_{\text{II}}(c_ib_j + c_jb_i)$$

There are six equations of this type. In addition, we have four linear equations which permit us to compute a_i in terms of b_i and c_i.

$$(3.2) \qquad n_i = n_{\mathrm{I}}\,(a_i + c_i) + n_{\mathrm{II}}\,(a_i + b_i) + n_{\mathrm{III}}a_i = na_i + n_{\mathrm{I}}c_i + n_{\mathrm{II}}b_i$$

We have now introduced ten of the $2^4 = 16$ available latent structure equations. Of the remaining, one is always:

$$n = n_{\mathrm{I}} + n_{\mathrm{II}} + n_{\mathrm{III}}$$

We have not yet utilized the four joint occurrences of three items at a time, nor n_{1234}.

In regard to the former, it is again useful to introduce them through cross products, this time of higher signature. From (2.1) and (2.5) we can adapt (3.1) to the following kind of equation:

$$(3.3) \qquad \begin{aligned}[12;4] = {} & n_{\mathrm{I}}\,(a_4 + c_4)\, n_{\mathrm{III}}a_4c_1c_2 \\ & + n_{\mathrm{II}}\,(a_4 + b_4)\, n_{\mathrm{III}}a_4b_1b_2 \\ & + n_{\mathrm{I}}\,(a_4 + c_4)\, n_{\mathrm{II}}\,(a_4 + b_4)\,(c_1 - b_1)\,(c_2 - b_2)\end{aligned}$$

It can be shown that for each triplet of items only one such equation contains independent information.

From (3.2) we can express a_4 in terms of b_4 and c_4, so that only terms in b_i and c_i appear in (3.3) for all the items involved. In other words, there is a simple way to construct a higher order from a lower order cross product. We just insert, every time that a latent class frequency appears in a formula, the corresponding latent marginal of the stratifying item.

The latent structure scheme of the present example shows that we have to find $(4\cdot 3) + 3 = 15$ latent parameters but have sixteen equations. In the mathematically perfect case, there has therefore to exist a condition between the manifest data. It is easily verified that the condition for $\lambda = 3$ and $m = 4$ is:

$$(3.4) \qquad \begin{vmatrix} n & n_1 & n_2 & n_{12} \\ n_3 & n_{13} & n_{23} & n_{123} \\ n_4 & n_{14} & n_{24} & n_{124} \\ n_{34} & n_{134} & n_{234} & n_{1234} \end{vmatrix} = 0$$

This means that we do not have to consider n_{1234} any further; it is dependent upon the other fifteen manifest frequencies. Therefore, it is not necessary to consider a sixteenth equation.

The problem now is to solve the ten equations (3.1) and (3.3) in

the ten unknowns b_i, c_i ($i = 1,2,3,4$) and n_{I} and n_{II}. Equations (3.1) can be made considerably simpler. We now introduce a new set of unknowns with the following transformation:

$$Y_i = b_i \sqrt{\frac{n_{\mathrm{II}}\,(n - n_{\mathrm{I}} - n_{\mathrm{II}})}{n\,(n - n_{\mathrm{I}})}}$$

(3.5)

$$Z_i = \sqrt{\frac{n_{\mathrm{I}}}{(n - n_{\mathrm{I}})\,n^2}}\,[c_i\,(n - n_{\mathrm{I}}) - b_i n_{\mathrm{II}}]$$

This transforms equation (3.1) into

(3.6) $$\frac{[ij]}{n^2} = Y_i Y_j + Z_i Z_j$$

For 4 test items this gives 6 equations in 8 unknowns.

It is not difficult to express the other 6 unknowns in terms of say, Y_4 and Z_4, from (3.6). But then we have to apply (3.5) to (3.3) to get 4 equations in b_4, c_4, n_{I} and n_{II}. This is the algebraic labor which has not yet been performed for the general case. But the general idea can be inducted from the case of $\lambda = 2$, and for specific cases of $\lambda > 2$, where the complete solution is relatively simple.

2. THE COMPLETE SOLUTION FOR $\lambda = 2$

Here the latent marginals are $a_i + b_i$ and a_i respectively. Equation (3.1) then becomes:

$$[ij] = n_{\mathrm{I}} n_{\mathrm{II}} b_i b_j$$

and the transformation (3.5) reduces to $Y_i = b_i \sqrt{\frac{n_{\mathrm{I}} n_{\mathrm{II}}}{n^2}}$ and $Z_i = 0$. Equations (3.6) are then $\frac{[ij]}{n^2} = Y_i Y_j$. They can be solved for any triplet of three items and give: $Y_i = S_i$ where $S_i = \sqrt{\frac{[ik]}{[kj]}\frac{[ij]}{n^2}}$. We thus have the difference between the two latent marginals expressed in terms of the latent class frequencies:

$$b_i = \frac{S_i}{\sqrt{\frac{n_{\mathrm{I}} n_{\mathrm{II}}}{n^2}}}$$

We take from (3.2) that

$$p_i = \frac{n_i}{n} = a_i + \frac{n_{\mathrm{I}}}{n} b_i$$

so that

$$a_i = p_i - S_i \sqrt{\frac{n_{\mathrm{I}}}{n_{\mathrm{II}}}} \tag{3.7a}$$

and

$$a_i + b_i = p_i + S_i \sqrt{\frac{n_{\mathrm{II}}}{n_{\mathrm{I}}}} \tag{3.7b}$$

Equation (3.3) in this case becomes:

$$[jk;i] = n_{\mathrm{I}} n_{\mathrm{II}} (a_i + b_i) a_i b_j b_k$$

Dividing by $[jk]$ we get:

$$\begin{aligned} \mu_i = \frac{[jk;i]}{[jk]} &= (a_i + b_i) a_i \\ &= \left(p_i + S_i \sqrt{\frac{n_{\mathrm{II}}}{n_{\mathrm{I}}}}\right)\left(p_i - S_i \sqrt{\frac{n_{\mathrm{I}}}{n_{\mathrm{II}}}}\right) \end{aligned} \tag{3.8}$$

μ_i incidentally is a "link" in the sense of Section II, paragraph 6. Equation (3.8) can be multiplied and the terms collected by the powers of $t = \sqrt{\frac{n_{\mathrm{I}}}{n_{\mathrm{II}}}}$. This finally gives the equation:

$$t^2 + Kt - 1 = 0 \tag{3.9a}$$

where

$$K = \frac{S_i}{p_i} - \frac{p_i}{S_i} + \frac{\mu_i}{S_i p_i} \tag{3.9b}$$

Equation (3.9a) permits us to compute the ratio $t = \sqrt{\frac{n_{\mathrm{I}}}{n_{\mathrm{II}}}}$, because K is known from manifest data (3.9b). By substituting t in (3.7a) and (3.7b) the whole latent structure becomes known. From equations (3.7) it can be seen that we would have to take the

positive sign for this root of t, because the model requires that $a_i + b_i > a_i$.

This sequence of computations suggests where the difficulties for a general solution arise and indicates in what kind of special cases they can be by-passed. Whenever it is possible to reduce equations (3.1) and (3.3) to fewer members, the algebra becomes more manageable. It becomes rather simple, if all members but one vanish. This is the case if there are only two distinct latent marginals for each item, even if $\lambda > 2$. It occurs in the "partially degenerate" structures discussed in paragraph 7 of Section II.

Fortunately, some of the structures which are most useful in practical work are of just this character. We shall now discuss three such restricted structures, showing (a) how their existence can be inferred from empirical data and (b) how the latent parameters can be computed if the conditions of reducibility are perfectly fulfilled. In Chapter 11, we will give concrete examples and computational procedures where these conditions are not quite met.[9]

3. A RESTRICTED QUADRATIC CASE

It is often convenient to think of a set of items as having a special kind of parabolic trace line. The probability of a positive answer is

[9] Two ways to solve the general latent structure equations of paragraph 1 might seem promising. They do not work, however, and the reason for their failure is instructive. One might first raise the question why equations (3.6) should not be used for more than 4 items and thus yield the basis for a simple solution. Now it is clear that for 6 items we get 15 equations, between which many relationships exist. We can form a variety of nonads between the cross products of $[ij]$ which will vanish because of (2.1). Therefore we do not have enough independent equations left to solve for 12 unknowns. With 5 items we have just 10 equations of type (3.6) and 10 unknowns. But unfortunately even here a rather complicated relationship exists between the ten cross products, as was shown in a letter by Professor L. J. Savage to the writer. Thus the problem is reduced to 4 items as developed in the present section.

But additional items might be used to take advantage of the "links" developed in Section II, paragraph 6. It was found, e.g., that

$$\mu_1 = \frac{\begin{vmatrix} [ik;1] & [il;1] \\ [jk;1] & [jl;1] \end{vmatrix}}{\begin{vmatrix} [ik] & [il] \\ [jk] & [jl] \end{vmatrix}} = (a_1 + c_1)(a_1 + b_1)\, a_1 \tag{3.10}$$

As the expression on the left is known from manifest data, this does indeed add an additional equation for the latent marginals of each of the 4 items. But (3.10) contains a_i on the right side and if we substitute from (3.2) this introduces n_{I} and n_{II}. Therefore, we need here again equations of type (3.3) for a final solution.

This is the reason, then, why our concrete examples (except for $\lambda = 2$) come from "partially degenerate" structures. These restrictions are put upon the latent marginals, so that the equations essentially boil down to forms similar to the case $\lambda = 2$. But the restrictions are such that the cases do not become "fully degenerate"—the latent parameters, as we shall see, can actually be computed.

greater, the higher a respondent is on the x-continuum. But the rate of increase is not uniform; the trace line is either convex or concave as shown in Figure 2.

Type 1 trace lines would be represented by quadratic parabolas with a minimum at point A; trace lines of type 2 would have a maximum at point B. The assumption in this simplified scheme is that the "concave" items all have their minimum at the same point A on the x-continuum and the "convex" ones their maximum at B. No people would be found on the continuum outside the line AB.

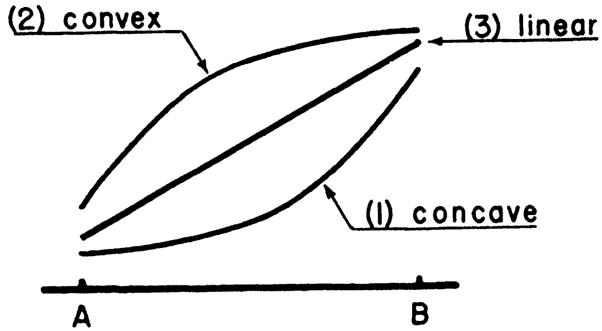


Figure 2

Such trace lines would represent a variety of situations well known in common test practices. Items may be worded positively or negatively as, e.g., in the following two statements:

(a) The British Army is fighting gallantly. Agree—Disagree.

(b) The British Army is not fighting gallantly. Agree—Disagree.

In order to express a pro-British attitude, which we define as a "positive" response, a respondent has to check "agree" after question *a* but "disagree" after question *b*. In the context of this test, question *a* is a "positively" worded question and question *b* is "negatively" worded. Suppose that respondents have a tendency to check "agree" rather than "disagree." This could be true especially for those who are in the middle range of an attitude continuum and who therefore have a harder time in making up their minds how to answer. If this is the case, question *a* might have a convex and question *b* a concave trace line.[10]

Another case in point is the handling of "don't knows." Suppose that a set of questions were asked of respondents, all of whom had

[10] It is important not to confuse a "positive" answer with the positive wording of a question. In the context of this test, "disagree" on question *b* is a "positive" answer to a "negatively" worded question.

made up their minds on each item, and that the trace lines of each item were reasonably linear, such as (3) in Figure 2. Now add people who say "don't know" in reply to some of the questions. Their position on the x-continuum will again be in the middle range between A and B. If we add the "don't knows" to the positive replies, then the trace line is likely to become convex; if we add them to the negative replies, the trace line is likely to become concave.

So much for the psychological background of the present model. Its equivalent latent structure has the advantage that it can be easily derived from the manifest data. It has, of course, three latent subclasses because the trace lines are quadratic. In addition it turns out that for each item two of the three latent marginals are alike. (The proof is not given here.) More specifically, the latent structure would look as follows:

TABLE 7

GENERAL LATENT STRUCTURE FOR RESTRICTED LATENT TRICHOTOMY

Latent class frequencies	*Latent marginals for items* 1	3	2	4
n_{I}	$a_1 + c_1$	$a_3 + c_3$	$a_2 + b_2$	$a_4 + b_4$
n_{II}	a_1	a_3	$a_2 + b_2$	$a_4 + b_4$
n_{III}	a_1	a_3	a_2	a_4

To facilitate the understanding of this structure, we shall use the following set of conventions according to whether an item is "concave" or "convex."

"concave"	*"convex"*
odd index numbers	*even* index numbers
index letters *plain*	index letters *with bar*
larger latent marginal: $a + c$	larger latent marginal: $a + b$

Table 7 is restricted to 4 items because the addition of items would add nothing to the basic algebra, as will be evident from the final formulae.

From equation (3.1) we can easily compute the manifest cross-product matrix which would be generated by the latent structure of Table 7.[11] (See Table 8.)

[11] This case is, of course, different from the "fully degenerate" structure discussed in Section II, paragraph 7. Here the equal latent marginals appear in different classes for "even" and for "odd" items.

TABLE 8

CROSS-PRODUCT MATRIX FOR RESTRICTED LATENT TRICHOTOMY

	1	*3*	*2*	*4*
1	—	$c_1c_3v_{\mathrm{I}}$	$b_2c_1n_{\mathrm{I}}n_{\mathrm{III}}$	$b_4c_1n_{\mathrm{I}}n_{\mathrm{III}}$
3	$c_1c_3v_{\mathrm{I}}$	—	$b_2c_3n_{\mathrm{I}}n_{\mathrm{III}}$	$b_4c_3n_{\mathrm{I}}n_{\mathrm{III}}$
2	$b_2c_1n_{\mathrm{I}}n_{\mathrm{III}}$	$b_2c_3n_{\mathrm{I}}n_{\mathrm{III}}$	—	$b_2b_4v_{\mathrm{III}}$
4	$b_4c_1n_{\mathrm{I}}n_{\mathrm{III}}$	$b_4c_3n_{\mathrm{I}}n_{\mathrm{III}}$	$b_2b_4v_{\mathrm{III}}$	—

The new symbols which appear here are introduced for the sake of brevity: $v_{\mathrm{I}} = n_{\mathrm{I}}\,(n - n_{\mathrm{I}})$ and $v_{\mathrm{III}} = n_{\mathrm{III}}\,(n - n_{\mathrm{III}})$.

This latent structure has 11 unknown parameters and leads to 16 equations. Therefore, five conditions have to be imposed upon the manifest data as conditions of reducibility. One of the conditions is equation (3.4). This condition applies because it holds for all cases where $m = 4$ and $\lambda = 3$.

Another condition can be seen from the latent structure scheme:

$$\text{(3.11a)} \qquad \begin{vmatrix} [12] & [14] \\ [32] & [34] \end{vmatrix} = 0$$

The existence of this condition permits the introduction of the symbol:

$$\text{(3.11b)} \qquad S_i = \sqrt{\frac{[i\bar{k}]\quad[ij]}{[j\bar{k}]\;n^2}}$$

where of items j and k one must be concave and the other convex. We see, e.g., that

$$\frac{[12]\,[13]}{[23]} = \frac{[14]\,[13]}{[34]}$$

Two more conditions can be derived from a further study of $[i\bar{k}] = n_{\mathrm{I}}n_{\mathrm{III}}b_{\bar{k}}c_i$. From the findings of Section II, paragraph 6, we know how to form $[i\bar{k};\bar{j}]$ or $[i\bar{k};l]$. We simply have to multiply any latent class frequency which appears in a formula by the corresponding latent marginals of the stratifying item. Thus we would find, e.g., that:

$$[12;3] = n_{\mathrm{I}}\,(a_3 + c_3)\,n_{\mathrm{III}}\,a_3b_2c_1$$

and

$$[14;3] = n_{\text{I}} (a_3 + c_3)\, n_{\text{III}}\, a_3 b_4 c_1$$

This makes it evident that there exists the condition that:

$$\frac{[12;3]}{[12]} = \frac{[14;3]}{[14]}.$$

A similar condition can be derived for "even numbered" stratifiers:

$$\frac{[j\bar{k};\bar{i}]}{[j\bar{k}]} = \frac{[l\bar{k};\bar{i}]}{[l\bar{k}]}$$

There then exists for each item, convex or concave, a unique ratio:

$$\eta_i = \frac{[l\bar{k};i]}{[l\bar{k}]} \tag{3.12}$$

which can be formed from the crossing of any two other items, as long as they are convex and concave respectively. This ratio, incidentally, should not be confused with any of the links, μ_i (see Section II, paragraph 6), which, in the present case of 3 latent subclasses, would be a ratio of two corresponding determinants of second order, with each element being an $[ij]$.

There exists still a fifth condition. Because of its more complex nature we shall come back to it later.

The solution now proceeds very much as in the case of $\lambda = 2$. From an inspection of the basic schemes in Table 8 we can multiply various cross products to get results of the following kind, e.g.,

$$\frac{[12]\,[13]}{[23]} = \frac{b_2 c_1^2 c_3 n_{\text{I}}^2 n_{\text{III}} (n - n_{\text{I}})}{b_2 c_3 n_{\text{I}} n_{\text{III}}} = c_1^2 n_{\text{I}} (n - n_{\text{I}})$$

These lead to the findings that

$$\begin{aligned} c_1 &= \frac{S_1}{\sqrt{\nu_{\text{I}} (1 - \nu_{\text{I}})}} \\ c_3 &= \frac{S_3}{\sqrt{\nu_{\text{I}} (1 - \nu_{\text{I}})}} \\ b_2 &= \frac{S_2}{\sqrt{\nu_{\text{III}} (1 - \nu_{\text{III}})}} \\ b_4 &= \frac{S_4}{\sqrt{\nu_{\text{III}} (1 - \nu_{\text{III}})}} \end{aligned} \tag{3.13}$$

where $\nu_{\text{I}} = \frac{n_{\text{I}}}{n}$ and $\nu_{\text{III}} = \frac{n_{\text{III}}}{n}$; these are the relative frequencies of those latent classes where each type of item has a marginal *different* from its marginals in the other two latent subclasses.

We now draw on the manifest marginals, p_i, of the items. Taking first a concave item we find from Tables 3 and 7 that

$$p_1 = a_1 + \nu_{\text{I}} c_1$$

Using (3.13) this gives

$$a_1 = p_1 - S_1 \sqrt{\frac{\nu_{\text{I}}}{1 - \nu_{\text{I}}}} \tag{3.14a}$$

$$a_1 + c_1 = p_1 + S_1 \sqrt{\frac{1 - \nu_{\text{I}}}{\nu_{\text{I}}}}$$

For convex items the marginals are:

$$p_2 = a_2 + (1 - \nu_{\text{III}}) b_2$$

Using (3.13) again this gives:

$$a_2 = p_2 - S_2 \sqrt{\frac{1 - \nu_{\text{III}}}{\nu_{\text{III}}}} \tag{3.14b}$$

$$a_2 + b_2 = p_2 + S_2 \sqrt{\frac{\nu_{\text{III}}}{1 - \nu_{\text{III}}}}$$

However, we do not know in terms of manifest data the values

$$t_{\text{I}} = \sqrt{\frac{\nu_{\text{I}}}{1 - \nu_{\text{I}}}} \tag{3.15}$$

and

$$t_{\text{III}} = \sqrt{\frac{1 - \nu_{\text{III}}}{\nu_{\text{III}}}}$$

To find them we remember that

$$\eta_1 = (a_1 + c_1) a_1$$

and

$$\eta_2 = (a_2 + b_2) a_2$$

where the coefficients on the left are formed according to (3.12) from manifest data.[12]

Using (3.14a) we find

$$\eta_1 = p_1^2 - S_1^2 + p_1 S_1 \left(\frac{1 - t_{\mathrm{I}}^2}{t_{\mathrm{I}}}\right)$$

Now we form the expression

$$K_{\mathrm{I}} = \frac{S_1}{p_1} - \frac{p_1}{S_1} + \frac{\eta_1}{p_1 S_1} \tag{3.16a}$$

and t_{I} is the root of the equation

$$t_{\mathrm{I}}^2 + K_{\mathrm{I}} t_{\mathrm{I}} - 1 = 0 \tag{3.17a}$$

In (3.11b) S_i is defined as a positive value. Therefore t_{I} has to be the positive root so that the obvious inequality $a_i + c_i > a_i$ is maintained.

In the same way we derive the equation

$$t_{\mathrm{III}}^2 + K_{\mathrm{III}} t_{\mathrm{III}} - 1 = 0 \tag{3.17b}$$

where

$$K_{\mathrm{III}} = \frac{S_2}{p_2} - \frac{p_2}{S_2} + \frac{\eta_2}{p_2 S_2} \tag{3.16b}$$

It can readily be demonstrated that t_{I} would not be different if instead of item 1 we had used the other concave item 3. In the same way, the convex items 2 and 4 are interchangeable for the formation of K_{III} and t_{III}. The inverse definitions of t_{I} and t_{III} in equation (3.15) however have to be kept carefully in mind.

The final solution then is given by equations (3.14a) and (3.14b) after the ratios of the latent class frequencies have been computed from equations (3.17a) and (3.17b). It will help in understanding the result, if one notices that K_{I} is formed from odd indices and

[12] It is now possible to name the fifth condition for reducibility, to wit:

$$\frac{t_{\mathrm{I}}^2}{t_{\mathrm{III}}^2} = \frac{[12]\,[34]}{[13]\,[24]} = \frac{[23]\,[14]}{[13]\,[24]}$$

The right-hand equality merely reproduces the second condition, equation (3.11a). The whole relationship can again be directly seen from the cross-product matrix of Table 8.

K_{III} from even indices. But the result is the same, whichever odd or whichever even item is used for the formation of K_I and K_{III} respectively.

4. THE RELATION BETWEEN TWO LATENT DICHOTOMIES

Suppose that two tests have each been found to be reducible to a latent dichotomy. Then by giving the two tests to the same sample it should be possible to study the relationship between the latent dichotomies. To simplify the problem, we shall assume that each test is "pure," i.e., the interrelationships between its items are fully accounted for by its own underlying continuum. But the position of a respondent on one continuum might well be related to his position on the other one. This would lead to the following latent structure. Items belonging to test A will be designated by odd indices; items from test B by even indices. People can be in four latent subclasses according to the following self-explanatory scheme:

		Latent dichotomy A +	Latent dichotomy A −	
Latent dichotomy B	+	n_I	n_{III}	n_B
	−	n_{II}	n_{IV}	
		n_A		n

We shall introduce 3 items for each test and their latent marginals will fit into the following scheme:

TABLE 9

LATENT STRUCTURE FOR TWO RELATED LATENT DICHOTOMIES

Latent class frequencies	*Latent marginals for items* 1	3	5	2	4	6
n_I	$a_1 + b_1$	$a_3 + b_3$	$a_5 + b_5$	$a_2 + b_2$	$a_4 + b_4$	$a_6 + b_6$
n_{II}	$a_1 + b_1$	$a_3 + b_3$	$a_5 + b_5$	a_2	a_4	a_6
n_{III}	a_1	a_3	a_5	$a_2 + b_2$	$a_4 + b_4$	$a_6 + b_6$
n_{IV}	a_1	a_3	a_5	a_2	a_4	a_6

For each item we have only two latent marginals since, by assumption, the probability of answering test A should not be related to position on the latent dichotomy for B, and inversely.

However we might well find that the two latent dichotomies themselves are related and that, therefore, $[AB] = n_I n_{IV} - n_{II} n_{III}$ will be greater or less than zero.

This then is a case with $m = 6$ and $\lambda = 4$ with restrictions upon the latent marginals but without restrictions on the latent class frequencies. There are sixteen latent parameters and 64 equations, so the conditions imposed upon the manifest data will be quite numerous. We shall follow up only those which are relevant for a solution in the mathematically perfect case.

The cross products between items which are both odd can easily be computed from formula (2.5a), e.g.,

$$\begin{aligned}[13] &= n_{\mathrm{I}}n_{\mathrm{II}}(a_1 + b_1 - a_1 - b_1)(a_3 + b_3 - a_3 - b_3) \\ &\quad + n_{\mathrm{I}}n_{\mathrm{III}}(a_1 + b_1 - a_1)(a_3 + b_3 - a_3) \\ &\quad + n_{\mathrm{I}}n_{\mathrm{IV}}b_1b_3 + n_{\mathrm{II}}n_{\mathrm{III}}b_1b_3 + n_{\mathrm{II}}n_{\mathrm{IV}}b_1b_3 \\ &\quad + n_{\mathrm{III}}n_{\mathrm{IV}}\cdot 0 \\ &= n_A(n - n_A)b_1b_3\end{aligned}$$

In the same way we find the cross products between even items, e.g.,

$$[24] = n_B(n - n_B)b_2b_4$$

The cross products between odd and even items is, e.g.,

$$\begin{aligned}[12] &= n_{\mathrm{I}}n_{\mathrm{II}}\cdot 0\cdot b_2 + n_{\mathrm{I}}n_{\mathrm{III}}b_1\cdot 0 + n_{\mathrm{I}}n_{\mathrm{IV}}b_1b_2 \\ &\quad - n_{\mathrm{II}}n_{\mathrm{III}}b_1b_2 + n_{\mathrm{II}}n_{\mathrm{IV}}b_1\cdot 0 + n_{\mathrm{III}}n_{\mathrm{IV}}\cdot 0\cdot b_2 \\ &= b_1b_2(n_{\mathrm{I}}n_{\mathrm{IV}} - n_{\mathrm{II}}n_{\mathrm{III}}) \\ &= [AB]\,b_1b_2\end{aligned}$$

The scheme of the cross-product matrix therefore is as follows:

TABLE 10

CROSS-PRODUCT MATRIX OF TWO RELATED LATENT DICHOTOMIES

	1	*3*	*5*	*2*	*4*	*6*
1	—	$b_1b_3v_A$	$b_1b_5v_A$	$b_1b_2\,[AB]$	$b_1b_4\,[AB]$	$b_1b_6\,[AB]$
3	$b_1b_3v_A$	—	$b_3b_5v_A$	$b_3b_2\,[AB]$	$b_3b_4\,[AB]$	$b_3b_6\,[AB]$
5	$b_1b_5v_A$	$b_3b_5v_A$	—	$b_5b_2\,[AB]$	$b_5b_4\,[AB]$	$b_5b_6\,[AB]$
2	$b_1b_2\,[AB]$	$b_3b_2\,[AB]$	$b_5b_2\,[AB]$	—	$b_2b_4v_B$	$b_2b_6v_B$
4	$b_1b_4\,[AB]$	$b_3b_4\,[AB]$	$b_5b_4\,[AB]$	$b_4b_2v_B$	—	$b_4b_6v_B$
6	$b_1b_6\,[AB]$	$b_3b_6\,[AB]$	$b_5b_6\,[AB]$	$b_6b_2v_B$	$b_6b_4v_B$	—

As usual, we designate

$$\begin{aligned}v_A &= n_A(n - n_A) \\ v_B &= n_B(n - n_B)\end{aligned}$$

The "odd" b_i values and v_A will be found by analyzing the odd items according to the formulae for $\lambda = 2$ developed in paragraph 2

of this section. In a corresponding way, we find the "even" b_E values and v_B. Once they are known, we have any number of odd-even cross products from which we can derive relationships like the following ones:

$$[AB] = \frac{[12]}{b_1 b_2} = \frac{[34]}{b_3 b_4} = \frac{[i\bar{k}]}{b_i b_E}$$

But from the treatment of the odd and even items separately we know that[13]

$$b_i = \frac{nS_i}{\sqrt{v_A}} \text{ and } b_E = \frac{nS_E}{\sqrt{v_B}}$$

for odd and even items respectively; the S-values are formed among the even and odd items separately. So we finally find that:

$$[AB] = \frac{[i\bar{k}]\sqrt{v_A v_B}}{n^2 S_i S_E}$$

From the knowledge of $[AB]$ and the formerly acquired knowledge of n_A and n_B we can of course reconstruct the whole latent fourfold table. It is worth mentioning that $\frac{[i\bar{k}]}{n^2 S_i S_E}$ is the Pearson r or the so-called phi-coefficient of the latent fourfold table, which appears here as a mere algebraic byproduct.

In paragraphs 5 and 6, Section V, this example will lead to a general discussion of structures with several underlying continua.

5. THE LATENT DISTANCE STRUCTURE

The last structure to be solved is reminiscent of items which form a so-called social distance scale. We shall assume $m = 3$ and $\lambda = 4$, and introduce a number of restrictions, which all can be derived from the following table:

TABLE 11

THE LATENT DISTANCE STRUCTURE

Latent class frequencies	*Latent marginals for items* *1*	*2*	*3*
n_{I}	$1 - d_1$	$1 - d_2$	$1 - d_3$
n_{II}	d_1	$1 - d_2$	$1 - d_3$
n_{III}	d_1	d_2	$1 - d_3$
n_{IV}	d_1	d_2	d_3

[13] For the meaning of S see paragraph 2 of this section.

Here the latent subclasses as well as the manifest items can be ordered in a unique way. If we assume the d-values to be less than .5, we can put it this way: in subclass I, the majority give a positive response to all three items; in subclass II only items 2 and 3 get a majority positive response; in subclass III only item 3 does. In subclass IV finally, the majority give a negative response to all three items.

The subclasses can be looked on as distances: each item is endorsed by a majority through a certain number of subclasses, and then "breaks" and falls to minority status. The value

$$x_j = \sum_{i=1}^{j} \frac{n_i}{n}$$

indicates how "far" through the latent subclasses the item j carries, and is an auxiliary latent parameter necessary for the computation of this structure.

An inspection of the scheme shows that there are only seven latent parameters; but there are still eight manifest response patterns, and therefore eight equations linking manifest data to latent parameters. One of these turns out to be dependent on the others, as can be shown as follows:

From (2.5a) we can derive the general relationship

$$[ij] = n^2 (1 - 2d_i)(1 - 2d_j)(x_i)(1 - x_j) \tag{3.18}$$

From an inspection of Table 11, we can see immediately that

$$x_j (1 - d_j) + (1 - x_j) d_j = p_j$$

and therefore

$$x_j = \frac{p_j - d_j}{1 - 2d_j} \tag{3.19}$$

Substituting equation (3.19) in equation (3.18) we obtain

$$[ij] = n^2 (p_i - d_i)(q_j - d_j) \tag{3.20}$$

where now the order of the indexes matters: $i < j$.

From equation (3.20) we can derive that

$$(3.21)\quad S_2^2 = \frac{[12]\,[23]}{n^2\,[13]} = (p_2 - d_2)\,(q_2 - d_2) = p_2q_2 - d_2\,(1 - d_2)$$

But by applying (2.5) we can also show that

$$(3.22)\qquad \eta_2 = \frac{[13;2]}{[13]} = d_2\,(1 - d_2)$$

Equations (3.21) and (3.22) together give us

$$(3.23)\qquad \eta_2 = p_2q_2 - S_2^2$$

This is a relationship between manifest constants which three items have to satisfy if a solution of the present problem is to be possible at all. If it is satisfied, then equation (3.23) permits the elimination of one of the eight equations. The one which we shall leave out is the one relating the highest joint frequency n_{123} to the latent parameters. Thus we are left with the task of solving the following seven equations for their seven unknown latent parameters:

$$x_j\,(1 - d_j) + (1 - x_j)\,d_j = \frac{n_j}{n} \qquad (j = 1, 2, 3)$$

$$x_j\,(1 - d_j)\,(1 - d_k) + (x_k - x_j)\,d_j\,(1 - d_k) + (1 - x_k)\,d_jd_k = \frac{n_{jk}}{n}$$

$$x_{IV} = 1$$

From equation (3.22) we immediately get d_2 as one root of the equation

$$d_2^2 - d_2 + \eta_2 = 0$$

From equation (3.20) we see that

$$d_1 = p_1 - \frac{[12]}{n^2\,(q_2 - d_2)}, \qquad d_3 = q_3 - \frac{[23]}{n^2\,(p_2 - d_2)}$$

Once the d-values are known, we get the auxiliary x-values from equation (3.19) and finally the latent class frequencies from the fact that $n(x_{II} - x_I) = n_{II}$, etc. In the next chapter we shall review all these solutions in the light of empirical data taken from Research Branch studies.

CHAPTER 11

THE INTERPRETATION AND COMPUTATION OF SOME LATENT STRUCTURES[1]

Introduction

IN THE next two sections a number of concrete examples, taken from Research Branch studies, are discussed in detail. Their purpose is threefold. First, they serve the function of illustration: they make more vivid and concrete the necessarily abstract deductions of the previous chapter. Secondly, they will afford an opportunity to introduce a number of new aspects of the general topic. Once the reader is familiar with the main argument, it is possible and profitable to follow up certain sidelines and details which were omitted in Sections II and III of Chapter 10. Finally, it will be seen that the actual computations involve new problems, because we deal in any concrete case with fallible data instead of exact theoretical relationships.

The computations have been presented in full so that the reader may easily check every step and form a judgment as to the character of the procedures and their practicality. No effort has been made here to present the most efficient computational procedures.[2] In many places we do not hesitate to interrupt the computational sequence in order to show some relationship to the general theory.

Thus, while the next two sections deal with empirical material, the main purpose still is to strengthen the presentation of the general theory. This is even more true for the last section, where a number of general test practices are related to the present approach.

[1] By Paul F. Lazarsfeld. The writer wishes to acknowledge the help his assistants have given him in the analysis of the examples upon which this chapter is based. Allen Barton has written a first report on most of the material included. Peter Rossi and Blanchard Lyon have added data and have carefully edited both Chapters 10 and 11.

[2] In connection with further work sponsored by the RAND Corporation, intensive study is being made of computational short-cuts, and of standardized forms which can be used by the ordinary statistical clerk.

SECTION IV

THE LATENT DICHOTOMY CASE

1. THE PROBLEM

So far we have proceeded on the assumption that all conditions are satisfied and that the number of latent classes is established as λ. As yet no attempt has been made to develop the sampling error for the theory; on the basis of sampling error alone, we can rarely expect these conditions to be met perfectly. One will have to make reasonable assumptions as to the number of latent classes with which one intends to work. In practice one will not start by trying to estimate the rank of a given joint frequency matrix; certain assumptions will be made as to what latent structure will fit the given data, and the work will consist in testing whether the assumptions are correct.

In this section all of our numerical examples will come from material which permits a reduction to two latent subclasses. This is a special case of a model where all trace lines are linear. In psychological terms, linear trace lines are expected for items in which nothing in the wording of the question would cause a sudden increase of positive replies at any point on the underlying x-continuum. The proportion of positive answers increases at a uniform rate as the respondents have higher and higher x-values. In the hypothetical example of Chapter 10, Figure 1, question 2 would be of the kind where a linear trace line might be expected. How to decide whether this expectation is borne out by the empirical data is the topic of the present section.

The trend of the argument is as follows: we compute the dichotomous latent structure which comes as near to the given data as possible. Then we derive from it a new set of response patterns which satisfy all the conditions corresponding to a model with linear trace lines. If the computed and the original frequencies are reasonably alike, our assumptions are justified. Otherwise, a different model has to be considered. If the deviations are too large, the first approach usually gives good leads as to how to proceed next. If the procedure is successful then the solution divides the given population into those who do and those who do not have a particular latent characteristic.

By starting with the simplest possible structure, we shall be able to illustrate all the necessary steps in the handling of empirical data

and in the interpretation of the results. After the reader has familiarized himself with these procedures, it will be easy for him to follow the manipulation of somewhat more complex structures to be discussed in the next section, since many of the procedures developed for the latent dichotomy will also apply in the subsequent cases.

From the solution given in Section III it can be seen that three items permit us to find all the parameters of a latent dichotomous structure, provided that the three items have positive cross products all through the system.

Let us summarize the necessary formulae for this case. We compute for one of the items, e.g., item 1, the values

$$S_1 = \sqrt{\frac{[12]\,[13]}{[23]\,n^2}} \qquad \text{and} \qquad \mu_1 = \frac{[23;1]}{[23]} \tag{4.1}$$

Then we compute the value

$$K = \frac{S_1}{p_1} - \frac{p_1}{S_1} + \frac{\mu_1}{S_1 p_1} \tag{4.2}$$

where p_1 is the proportion of "positive" replies on item 1 for the whole sample. (The same K-value would have resulted had we started with item 2 or 3.) We then take the positive root of the equation

$$t^2 + Kt - 1 = 0 \tag{4.3}$$

This root is $t = \sqrt{\frac{n_{\mathrm{I}}}{n_{\mathrm{II}}}}$, where n_{I} and n_{II} are the two latent class frequencies, and $n_{\mathrm{I}} + n_{\mathrm{II}} = n$. Thus from t we compute the two latent class frequencies.

We find that

$$t = \frac{-K}{2} + \sqrt{\frac{K^2 + 4}{4}} \tag{4.4}$$

and

$$\nu_{\mathrm{I}} = \frac{n_{\mathrm{I}}}{n} = \frac{t^2}{1 + t^2} \tag{4.5}$$

It is also helpful to remember that $K = \dfrac{n_{\mathrm{II}} - n_{\mathrm{I}}}{\sqrt{n_{\mathrm{I}} n_{\mathrm{II}}}}$.

Finally we compute the latent marginals from:

$$(4.6)\qquad \begin{aligned} p_{\mathrm{I}j} &= p_j + \frac{S_j}{t} \\ p_{\mathrm{II}j} &= p_j - S_j t \end{aligned}$$

We have dealt with the simple three-item case. Where four items are used, the computations are somewhat more involved. For the four-item case, the latent structure is given by the following scheme:

TABLE 1

THE LATENT STRUCTURE OF A LATENT DICHOTOMY

Latent class frequencies	*Latent marginals for items* 1	2	3	4
n_{I}	$a_1 + b_1$	$a_2 + b_2$	$a_3 + b_3$	$a_4 + b_4$
n_{II}	a_1	a_2	a_3	a_4

This structure has 10 latent parameters, and 16 manifest response patterns. Therefore, 6 conditions will have to be satisfied between the 16 manifest joint occurrences (they were summarized in Section II, paragraph 7). For the present purpose these conditions will be presented in a slightly different way.

Four items can be combined into 4 "triplets." Besides the combination {234}, item 1 can be combined with items 2 and 3, or with 2 and 4, or with 3 and 4. This gives three values for S_1, one for each triplet used. If we take the three triplets involving item 1 we would have

$$(4.7a)\qquad \begin{aligned} S_1' &= \sqrt{\frac{[1\,2]\,[1\,3]}{n^2\,[2\,3]}} \\ S_1'' &= \sqrt{\frac{[1\,2]\,[1\,4]}{n^2\,[2\,4]}} \\ S_1''' &= \sqrt{\frac{[1\,3]\,[1\,4]}{n^2\,[3\,4]}} \end{aligned}$$

In the perfect case

$$(4.7b)\qquad S_1' = S_1'' = S_1'''$$

because $\frac{[13]}{[23]} = \frac{[14]}{[24]}$ and $\frac{[12]}{[23]} = \frac{[14]}{[34]}$ according to proposition (a) in Section II, paragraph 6. In other words, S_j *is a value attached to item* j *irrespective of which other two items are used for numerical computations,* as long as the whole system is perfectly reducible to two homogeneous systems.

There are also three "links" for each item. They also are alike in the perfect case because

$$\frac{[23;1]}{[23]} = \frac{[24;1]}{[24]} = \frac{[34;1]}{[34]} \tag{4.7c}$$

Therefore μ_j *is ideally also a value attached to item* j *irrespective of what other items have been used* in the formation of this parameter.

This necessarily leads to the finding that, under conditions of perfect reducibility, K should have the same value for any triplet of items in the system, since, as can be seen from (4.2), in the formation of K only S_j, μ_j, and p_j are involved and the latter three have a unique value for item j, according to (4.7b) and (4.7c) above. The reader should keep the following distinction in mind: it is an algebraic fact that, for a given triplet of items, e.g., $\{123\}$, K computed from item 1 is identical with K computed from item 2, and with K computed from item 3, regardless of whether the entire system meets the conditions of reducibility. However, *only under conditions of reducibility* is the value of K constant for all different possible triplets in the system, e.g., K based on $\{123\}$ and K based on $\{124\}$ are identical.

Thus, with empirical data, the values of S, μ, and K will vary from one triplet to the next. Therefore, if a latent dichotomy is to be fitted, adjustments have to be made. It is these adjustments which form the main task in actual computations.

2. THE FITTED CROSS-PRODUCT MATRIX

We now introduce a numerical example in order to show the computations step by step. We will use four items from a questionnaire on attitude toward the Army. The items are:

1. In general, how well do you think the Army is run? Positive answer: *Very well* or *Pretty well*	64.1%
2. Do you think when you are discharged you will go back to civilian life with a favorable or unfavorable attitude toward the Army? Positive answer: *Very* or *Fairly favorable*	37.4

3. In general, do you feel you yourself have gotten a square deal from the Army?
   Positive answer: *Yes, in most ways I have* 30.0

4. Do you feel that the Army is trying its best to look out for the welfare of enlisted men?
   Positive answer: *Yes, it is trying its best* 25.4

These four dichotomies give us sixteen pieces of manifest data (Table 2) which can be expressed either in terms of positive frequencies or as response patterns of positive and negative responses.

TABLE 2

MANIFEST DATA OF FOUR ITEMS ON ATTITUDE TOWARD THE ARMY

(a) *As positive frequencies*			(b) *As response patterns* *1*	*2*	*3*	*4*	
n	= 1,000 (sample size)		+	+	+	+	75
n_1	= 641	marginals	+	+	+	−	69
n_2	= 374	marginals	+	+	−	+	55
n_3	= 300	marginals	+	−	+	+	42
n_4	= 254	marginals	−	+	+	+	3
n_{12}	= 295	joint frequencies	+	+	−	−	96
n_{13}	= 246	joint frequencies	+	−	+	−	60
n_{14}	= 217	joint frequencies	+	−	−	+	45
n_{23}	= 163	joint frequencies	−	+	+	−	16
n_{24}	= 141	joint frequencies	−	+	−	+	8
n_{34}	= 130	joint frequencies	−	−	+	+	10
n_{123}	= 144	triple joint frequencies	+	−	−	−	199
n_{124}	= 130	triple joint frequencies	−	+	−	−	52
n_{134}	= 117	triple joint frequencies	−	−	+	−	25
n_{234}	= 78	triple joint frequencies	−	−	−	+	16
n_{1234}	= 75		−	−	−	−	229

Indeed, we can get the positive frequencies from the response patterns and vice versa. It might be noted that n_{1234} is identical with the response pattern $(+ + + +)$.

We now form a cross-product matrix by computing the basic determinants $[ij]$. These are simply the cross product of a fourfold table showing the combinations of positive and negative occurrences of two items:

		Item i +	*Item i* −	
Item j	+	a or n_{ij}	b	$n_j = a + b$
	−	c	d	

$n_i = a + c \qquad n = a + b + c + d$

$[ij] = ad - bc = n \cdot n_{ij} - n_i n_j$

Table 3 shows the cross-product matrix for these four items. The cross products have been standardized by dividing each one by the square of the sample size.

TABLE 3

EMPIRICAL CROSS-PRODUCT MATRIX: $[ij]/n^2$

	1	*2*	*3*	*4*
1	—	.0553	.0537	.0542
2	.0553	—	.0508	.0460
3	.0537	.0508	—	.0538
4	.0542	.0460	.0538	—

From the cross products we can then form the S-values according to formula (4.7a). As we mentioned before, we get three slightly different values for every item, according to which triplet we consider. The three forms of S_1 from (4.7a) have respectively the values:

$$S_1' = \sqrt{\frac{[12]\,[13]}{n^2\,[23]}} = \sqrt{\frac{55{,}266 \times 53{,}700}{(1000)^2\,50{,}800}} = .2417 \text{ from } \{123\}$$

$$S_1'' = \sqrt{\frac{[12]\,[14]}{n^2\,[24]}} = \sqrt{\frac{55{,}266 \times 54{,}186}{(1000)^2\,46{,}004}} = .2551 \text{ from } \{124\}$$

$$S_1''' = \sqrt{\frac{[13]\,[14]}{n^2\,[34]}} = \sqrt{\frac{53{,}700 \times 54{,}186}{(1000)^2\,53{,}800}} = .2326 \text{ from } \{134\}$$

In the perfect case these three values would be alike. With fallible data a "true" value for S will have to be estimated and in the present procedure the "true" value is taken as the geometric mean of the three actual values. The geometric mean is chosen because of the great ease with which it can be computed using the logarithms of the cross products. We translate the *cross-product*

matrix into its logarithmic form as in Table 4, which is derived from Table 3 and a table of logarithms. Note that the cross products in Table 4 are *not* standardized by division by n^2.

TABLE 4

LOGARITHMS OF EMPIRICAL CROSS-PRODUCT MATRIX FROM TABLE 2: log $[ij]$

	1	*2*	*3*	*4*
1	——	**4.742458**	**4.729974**	**4.733887**
2	4.742458	——	*4.705864*	*4.662796*
3	4.729974	4.705864	——	*4.730782*
4	4.733887	4.662796	4 730782	——

From Table 4 we can compute directly the logarithm of the "true value" of S_j. It can be easily verified that S_j can be computed from the following formula:[3]

$$\log S_j = \frac{1}{m-1} \sum \log [ij] - \frac{1}{(m-1)(m-2)} \sum \sum \log [ik] - \log n \tag{4.8}$$

where m is the number of items in the matrix; in our example, $m = 4$. The first summation in the formula is taken over all cross products involving j. If $j = 1$ this would be the sum of all the figures in the first row of Table 4. A cross-product matrix can, of course, always be so arranged that any item is in the first row. The double sum covers all the items to the right of the diagonal which are not in the first row. Formula (4.8) can easily be remembered by noticing that the two sums on the right are formed over the boldface figures and the italicized figures, respectively, as indicated in Table 4, as long as item j is thought of as determining the first row. The numerical results (from 4.8) are, in the present example:

$$\begin{aligned}
\log S_1 &= .385533 - 1 \\
\log S_2 &= .337932 - 1 \\
\log S_3 &= .365683 - 1 \\
\log S_4 &= .346106 - 1
\end{aligned}$$

The *final values of the adjusted* S_j are therefore

[3] In order to avoid cumbersome symbolism, the summations are explained in the text.

$$S_1 = .24296 \qquad S_2 = .21774$$
$$S_3 = .23210 \qquad S_4 = .22187$$

From (4.8) we can now compute adjusted or *"fitted" cross products.* In the case of perfect reducibility, the relationship holds that

$$\frac{[ij]}{n^2} = S_i S_j \tag{4.9a}$$

and

$$\log\,[ij] = \log S_i + \log S_j + 2 \log n \tag{4.9b}$$

In the case of imperfect data we can introduce "fitted" cross products $[ij]$ by applying (4.9) to the adjusted S-values. Thus, for example, the fitted $\frac{[14]}{n^2} = S_1 S_4 = .24296 \times .22187 = .0539$, which is slightly different from the original value .0542. The actual computations are usually made with logarithms.

The numerical values of the actual and the "fitted" cross products (divided by the square of the sample) are given in Table 5. The fitted values, of course, satisfy perfectly the condition that tetrads vanish.

TABLE 5

COMPARISON OF ACTUAL AND FITTED CROSS-PRODUCT MATRICES DERIVED FROM TABLE 2

	1	*2*	*3*	*4*
1	Actual Fitted Difference	.0553 (.0529) +.0024	.0537 (.0564) −.0027	.0542 (.0539) +.0003
2	.0553 (.0529) +.0024	—	.0508 (.0505) +.0003	.0460 (.0483) −.0023
3	.0537 (.0564) −.0027	.0508 (.0505) +.0003	—	.0538 (.0515) +.0023
4	.0542 (.0539) +.0003	.0460 (.0483) −.0023	.0538 (.0515) +.0023	—

Comparing the actual and the fitted data gives a general idea how hopeful it would be to try a latent dichotomy structure. If at this point there is wide divergence, we would know that we are on the wrong track. If the differences are small, we will proceed with the analysis.[4]

It should be made clear that in the absence of an exact sampling theory this judgment can be based only on common sense and has no precise justification. Two general considerations, however, give some assistance at this point. First of all, the deviations can be of two kinds: they may be of approximately equal size, scattered all over the matrix, or they may be concentrated within a small subgroup of items. If, for example, the cross products among a few items have positive deviations in all combinations of these items, then this is a strong indication that the conditions of homogeneity for a "pure" test are not fulfilled (see Section I, paragraph 2, especially equation 1.1). There is likely to exist a "group factor" which ties these items together beyond what can be accounted for by one underlying continuum. In such a case another latent structure has to be postulated and its existence tested, as, for example, the latent turnover developed in Chapter 10, Section III. In the present example, no triplet has all of its three deviations in the same direction and no group factors seem to be present.

In this case the judgment can be guided by a second consideration, namely that the deviations between actual and fitted cross-product matrices have a direct meaning. They are the deviation of the actual joint frequency of the two items from the joint frequency which would fit the conditions perfectly, the deviation being expressed as a proportion of n, the sample size. Thus the difference between the observed $[12]/n^2$ and the fitted $[12]/n^2$ is .0024. This means that n_{12} is greater than it should be for a perfect fit by .0024 of the sample size. Since $n = 1000$, n_{12} is 2.4 cases too large to meet the tetrad condition perfectly. Experience shows that this is a difference which would not turn out to be very disturbing in the eventual fit of the latent structure to the manifest data.

3. THE NUMERICAL COMPUTATION OF THE LATENT STRUCTURE

In order to compute the latent marginals and class frequencies through formulae (4.5) and (4.6) we must now get the value

[4] The above procedure, while developed independently, was later found to be very similar to one developed for the Spearman general factor theory. See L. L. Thurstone *Multiple-Factor Analysis* (University of Chicago Press, 1947), p. 277.

$$K = \frac{S_j}{p_j} - \frac{p_j}{S_j} + \frac{\mu_j}{p_j S_j} \tag{4.10}$$

It will be useful to carry out the computation of K for one triplet, e.g., $\{123\}$, in all detail, since this parameter plays a basic role in the whole procedure. This will also provide an occasion to review once more the formation of the other parameters, namely S and μ. (In working with more than 3 items, one would not compute all the separate K-values; one would insert in 4.10 the adjusted S-value computed by 4.8. The separate computations have been included here to give the reader complete information on one triplet.)

Our first job is to compute the p-, S- and μ-values:

$$p_1 = \frac{n_1}{n} = .641, \qquad p_2 = \frac{n_2}{n} = .374, \qquad p_3 = \frac{n_3}{n} = .300$$

For the S-values we use the cross products $[ij]$. Table 3 already contains the computed cross products (divided by the square of the sample size); the computation of these basic determinants is shown in paragraph 2.

Therefore

$$S_1 = \frac{1}{n}\sqrt{\frac{[12]\,[13]}{[23]}} = \frac{1}{1000}\sqrt{\frac{55{,}266 \times 53{,}700}{50{,}800}} = .241704$$

$$S_2 = \frac{1}{n}\sqrt{\frac{[12]\,[23]}{[13]}} = \frac{1}{1000}\sqrt{\frac{55{,}266 \times 50{,}800}{53{,}700}} = .228651$$

$$S_3 = \frac{1}{n}\sqrt{\frac{[13]\,[23]}{[12]}} = \frac{1}{1000}\sqrt{\frac{53{,}700 \times 50{,}800}{55{,}266}} = .222172$$

Finally, we require the μ-values, e.g.,

$$\mu_1 - \frac{[23;1]}{[23]}$$

We already have [23]. The quantity [23;1] actually is the value of [23] for *only those who are positive on item 1:* in other words, the cross product of the following fourfold table:

Positive on Item 1

		Item 2 +	Item 2 −	
Item 3	+	n_{123} = 144	102	$n_{13} = 246$
	−	151	244	= 395
		$n_{12} = 295$	346	$n_1 = 641$

$$\begin{aligned}[23;1] &= n_1 n_{123} - n_{12} n_{13} \\ &= 641 \times 144 - 295 \times 246 = 19{,}734\end{aligned}$$

In the same way

$$[13;2] = n_2 n_{123} - n_{12} n_{23} = 374 \times 144 - 295 \times 163 = 5771$$
$$[12;3] = n_3 n_{123} - n_{13} n_{23} = 300 \times 144 - 246 \times 163 = 3102$$

We can now easily find the μ-values for triplet $\{123\}$:

$$\mu_1 = \frac{[23;1]}{[23]} = \frac{19{,}734}{50{,}800} = .388465$$

$$\mu_2 = \frac{[13;2]}{[13]} = \frac{5771}{53{,}700} = .107467$$

$$\mu_3 = \frac{[12;3]}{[12]} = \frac{3102}{55{,}266} = .0561285$$

We can now compute the K-value for this triplet using any of these three sets of values of p, S, and μ. It is generally a good idea to compute it all three ways as a check against computing mistakes.

Using p-, S-, and μ-values based on item 1:

$$K = \frac{.241704}{.641} - \frac{.641}{.241704} + \frac{.388465}{.241704 \times .641} = .232391$$

Using the values based on item 2:

$$K = \frac{.228651}{.374} - \frac{.374}{.228651} + \frac{.107467}{.228651 \times .374} = .232382$$

Using the values based on item 3:

$$K = \frac{.222172}{.300} - \frac{.300}{.222172} + \frac{.0561285}{.222172 \times .300} = .232386$$

The values disagree slightly due to the rounding off of errors; the difference is due purely to the mechanics of computation. For the first triplet, therefore, we can say that $K = .23239$.

In a perfectly reducible case of more than 3 items, K would be the same for every triplet. Among other conditions, (4.7c) has to be satisfied to give us a unique μ_j-value for each item. With fallible data this will in general not be the case. For item 1, e.g., we obtain

$$\frac{[23;1]}{[23]} = .38846$$

$$\frac{[24;1]}{[24]} = .41985$$

$$\frac{[34;1]}{[34]} = .40177$$

Thus, for each triplet we will find a slightly different K, even if we use the adjusted S-values as developed in (4.8). The adjustment at this point cannot be condensed into a simple formula like (4.8). Considerable experimentation will be necessary to find out how much different adjustment procedures affect the final result. If many items have to be considered, a sampling of triplets will be necessary to make the whole procedure manageable. Investigations in this direction are under way.

For the present example it is simplest to compute the proper K for each triplet by using (4.10). Then, by an application of (4.4) and (4.5) to this value of K, we may obtain a value of n_{I} for each triplet, as follows:

Triplet	*K-value*	n_{I}
{123}	.23239	442.3
{124}	.45290	389.6
{134}	.30167	425.4
{234}	.23862	440.8

The average n_{I} is 424.5. It would also have been possible to average the K-values and to compute one n_{I} only. This would have given the result $n_{\text{I}} = 424.3$, which scarcely differs from the other value.

Rather than compute n_{I} by computing t using formulae (4.4) and (4.5), a nomograph relating K and t directly may be employed and the values of t corresponding to the computed value of K can be read off directly. This nomograph is reproduced as Figure 1.

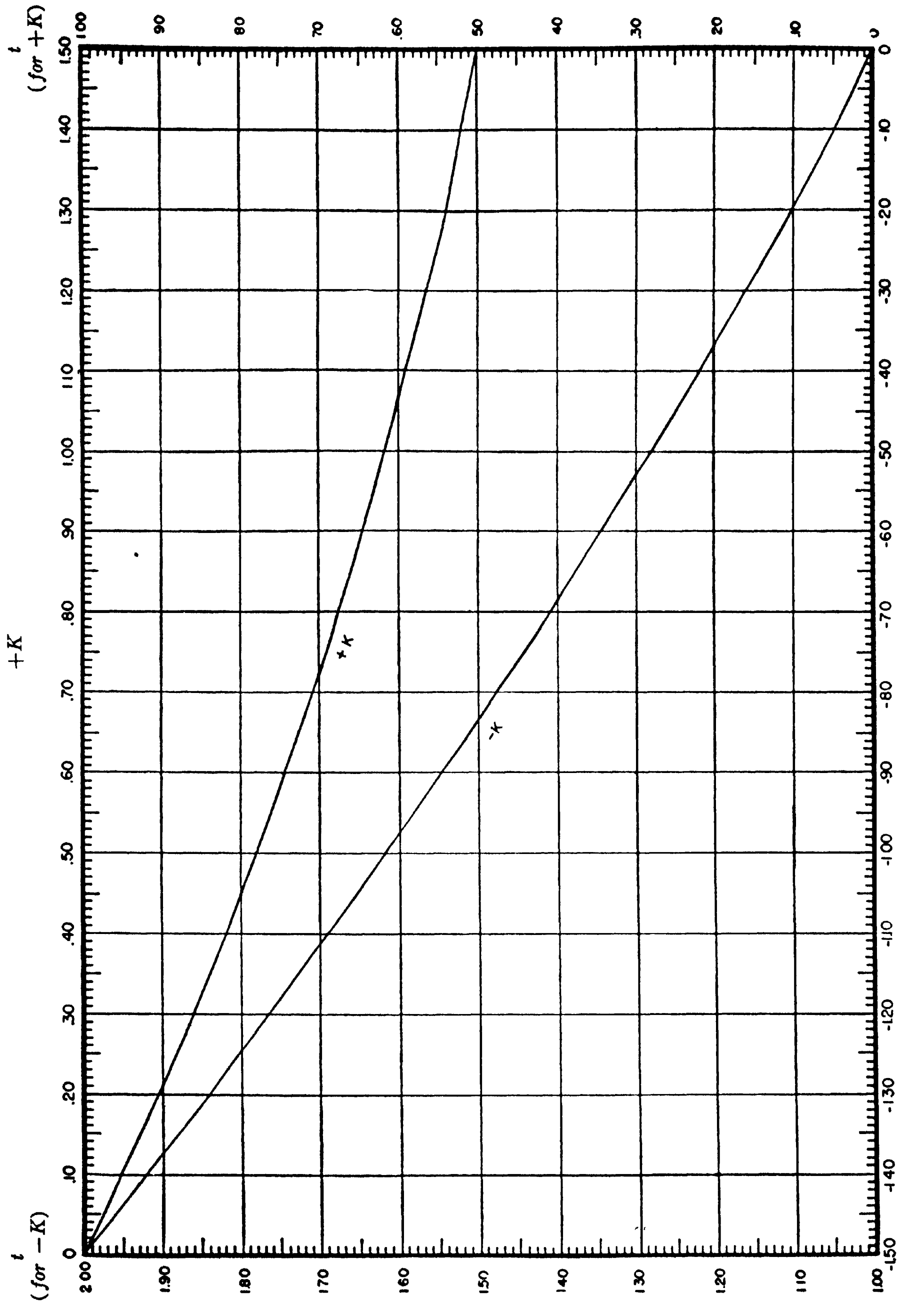


Figure 1. Graph of K and t. The nomograph relates the values of K to the corresponding values of t. Note that there are two sets of scales. Along the upper horizontal edge are placed positive values of K and the line marked "$+K$" relates each value of positive K to the values of t on the right-hand vertical scale. The value of t for positive values of K is found by using this set of scales.

The other set of scales is for negative values of K. Along the lower horizontal edge are placed negative values of K and the line marked "$-K$" relates this scale to the left-hand vertical scale of t.

Once the latent class frequencies are computed, the latent marginals follow easily from formula (4.6). We have already computed the adjusted S-values in paragraph 2, because they were needed for the fitted cross-product matrix. From (4.4) and using the average K-value above, we may find that $t = \sqrt{\frac{n_I}{n_{II}}} = .85847$. The latent marginals therefore are:

$$
\begin{aligned}
p_{I1} &= p_1 + \frac{S_1}{t} = .641 + \frac{.24296}{.85847} = .92402 \\
p_{II1} &= p_1 - S_1 t = .641 - .24296 \times .85847 = .43243 \\
p_{I2} &= p_2 + \frac{S_2}{t} = .374 + \frac{.21774}{.85847} = .62764 \\
p_{II2} &= p_2 - S_2 t = .374 - .21774 \times .85847 = .18708 \\
p_{I3} &= p_3 + \frac{S_3}{t} = .300 + \frac{.23210}{.85847} = .57036 \\
p_{II3} &= p_3 - S_3 t = .300 - .23210 \times .85847 = .10075 \\
p_{I4} &= p_4 + \frac{S_4}{t} = .254 + \frac{.22187}{.85847} = .51245 \\
p_{II4} &= p_4 - S_4 t = .254 - .22187 \times .85847 = .06353
\end{aligned}
$$

We now have computed the complete latent structure for our four items, assuming that $\lambda = 2$. It can be presented in summary form:

TABLE 6

COMPUTED LATENT STRUCTURE FOR ATTITUDE TOWARD ARMY

Latent class frequencies	*Latent marginals of items* 1	2	3	4
$n_I = 424.3$	.9240	.6276	.5704	.5125
$n_{II} = 575.7$	.4324	.1871	.1008	.0635

4. THE COMPUTATION OF THE FITTED RESPONSE PATTERNS

We can now compute the set of sixteen response patterns which would correspond exactly to this latent structure. We have two latent subclasses which are both homogeneous. Therefore the number of people giving a certain response pattern can be computed by multiplying the appropriate pattern of positive and negative latent marginals *in each subclass*. The frequencies are then added up to

get the number of respondents for a given response pattern for the whole sample.

The pattern (+ − + −) would be, e.g., given by the following number of respondents:

$$n_{I}p_{I1}q_{I2}p_{I3}q_{I4} + n_{II}p_{II1}q_{II2}p_{II3}q_{II4} \qquad \text{(where } q = 1 - p\text{)}$$

In Table 7 the first column gives the response pattern; the second and third columns report the contribution of each latent class to this pattern respectively. The fourth column reports the sum of the preceding two: it gives the number of people who would answer in this way, *if the latent structure were perfectly satisfied.* The fifth column tells how many people actually answered in each pattern.

The discrepancies between columns IV and V are a measure of the adjustment which was necessary to find a solution for the given problem. We compare our actual and a "fitted" distribution of response patterns, just as we would if we had to fit a normal distribution curve to a given set of data. The fitted response frequencies can be computed if we have an S-value for each item and a K-value $\left(\text{and therefore a } t = \sqrt{\frac{n_{I}}{n_{II}}}\right)$ for the whole system. It is always

TABLE 7

FITTED AND ACTUAL RESPONSE PATTERN FREQUENCIES FOR ATTITUDE TOWARD ARMY

I *Response patterns*	*II* *Latent class I*	*III* *Latent class II*	*IV* *Total fitted*	*V* *Actual*	*VI* *Ratio* $\frac{\text{Col. II}}{\text{Col IV}}$
+ + + +	71.9	.3	72.2	75	.996
+ − + +	42.7	1.3	44.0	42	.971
+ + − +	54.2	2.6	56.8	55	.953
+ + + −	68.4	4.4	72.8	69	.940
− + + +	5.9	.4	6.3	3	.938
+ − − +	32.2	11.5	43.7	45	.736
+ − + −	40.6	19.1	59.7	60	.680
− − + +	3.5	1.7	5.2	10	.674
+ + − −	51.6	39.2	90.8	96	.568
− + − +	4.5	3.5	8.0	8	.561
− + + −	5.6	5.8	11.4	16	.494
+ − − −	30.6	170.4	201.0	199	.152
− − − +	2.6	15.2	17.8	16	.148
− − + −	3.3	25.1	28.4	25	.117
− + − −	4.2	51.5	55.7	52	.076
− − − −	2.5	223.7	226.2	229	.011
Totals	424.3	575.7	1000.0	1000	

possible to compute such average auxiliary parameters. Whether the final fit to which they lead is acceptable can be decided only in the light of the problem at hand. We shall come back to this point repeatedly.

Professor Frederick Mosteller has suggested that the goodness of fit could be tested by using χ^2 with 2^m degrees of freedom reduced by the number of latent parameters. Experiments are under way to see whether this idea would be borne out in hypothetical models based on random numbers.

It is worth noticing that in the procedure developed two kinds of adjustments have been used. The joint frequencies of two and of three items were numerically adjusted so as to conform to the conditions of reducibility. This was done by computing adjusted S-values which involved joint frequencies between pairs of items only; and by computing an average K, involving triplets of items. The joint positive occurrence of all four items, however, did not enter the computations at all. They were disregarded in the computation of the latent parameters. Only at the end when the ultimate frequencies were computed did they enter the picture. However, we see that there is not too great a discrepancy between the number of people who actually give a positive reply to all four questions (n_{1234}) and the adjusted frequency $\sum_j n_j p_{j1} p_{j2} p_{j3} p_{j4}$.

5. THE ORDERING OF RESPONSE PATTERNS AND THE CLASSIFICATION OF RESPONDENTS

We now turn our attention to the final column in Table 7. Each response pattern, as it were, recruits its respondents from the two latent subclasses. There are, e.g., 16 people who answer $(- + + -)$ and 60 who answer $(+ - + -)$. Of the former, 49 per cent come from the positive subclass; of the latter, 68 per cent are thus recruited. We can use this ratio as an *ordering device for the response patterns.* Table 7 has been arranged accordingly. The latent dichotomy structure thus permits a unique ordering of response patterns.

It is possible to go a step further in the interpretation of this ordering ratio. *It is the probability that a person giving this response pattern comes from class I rather than class II.* This can easily be seen if one remembers how the two elements of this ratio were derived. We take, as an example, the people with response pattern $(- + + -)$. In class I the probability of giving this response was

$q_{11}p_{12}p_{13}q_{14}$; in class II, it was $q_{111}p_{112}p_{113}q_{114}$. There is no other place from which these respondents can be recruited. Therefore the reasoning of a posteriori probability applies here. The probability that such a respondent comes from the first class is the product of the probability of being in class I, times the probability of giving this answer *if* one is in this class. Now the probability of being in class I is $\frac{n_{\text{I}}}{n}$; therefore, the probability of being in class I and giving response pattern $(- + + -)$ is $\frac{n_{\text{I}}q_{11}p_{12}p_{13}q_{14}}{n}$. The alternative probability for class II is $\frac{n_{\text{II}}q_{111}p_{112}p_{113}q_{114}}{n}$. The a posteriori probability therefore that a response pattern $(- + + -)$ comes from class I is:

$$\frac{n_{\text{I}}q_{11}p_{12}p_{13}q_{14}}{n_{\text{I}}q_{11}p_{12}p_{13}q_{14} + n_{\text{II}}q_{111}p_{112}p_{113}q_{114}}$$

This is exactly how the ratio under discussion was formed.

Table 7, therefore, not only orders the response patterns; it also permits us to classify respondents into the two subclasses. This cannot be done with perfect certainty, but only with a specified probability attached to the judgment. We see, for example, that 99.6 per cent of the people who answer $(+ + + +)$ belong in class I and 98.9 per cent of the $(- - - -)$ respondents belong in class II (because $1 - .011 = .989$). These two types of respondents will therefore be attributed to latent classes with a very small degree of error. But when it comes to the 96 people who answer $(+ + - -)$ the misclassification we are forced to make is much greater. Of this group, 56.8 per cent come from class I and 43.2 per cent from class II. It will still be the best judgment to put them into class I; but we know that 41 cases of the 96 will be misclassified.

If one looks at the four items of Table 7 as a test, then a *measure of precision* can be derived from the preceding considerations. We may classify each respondent into that latent class where he most probably belongs; but we know that we cannot classify all respondents correctly. Comparing columns II and III of Table 5 row by row, we shall have to classify the respondents in the top 10 rows as "latent positive"; but we see that this will be correct for 375.5 cases only (those in column II). The bottom 6 groups of re-

spondents will have to be called "latent negative"; but again we shall be correct only for the 491.7 cases in column III. The "precision" of the test is therefore only $\frac{375.5 + 491.7}{1{,}000} = 86.7$ per cent.

It is important not to confuse *precision* and *fit*. The latter corresponds to a comparison of columns IV and V in Table 7. Discrepancies are due partly to sampling error and partly to the fact that the assumed latent structure is not quite the one from which the actual data were generated. A certain amount of imprecision, however, will exist even if the data match the latent structure perfectly. The attribution of a respondent to a latent class is a probability judgment based on his manifest response pattern. It can therefore never be made with complete certainty.

Furthermore this attribution might change as the amount of manifest information increases. Assume for a moment that only the first three items of our example had been available and concentrate on the people answering (+ + −) and (+ − −). From Table 7 it can be seen that 151 would have been in the first group and 244 in the second. Their ordering ratio would have been .717 and .257 respectively. Now we add the fourth item. This splits each group in the following way:

	Ordering ratio for first 3 items	*Ordering ratio for all four items*	*Number of cases*
+ + − +	.717	.953	55
+ + − −		.568	96
+ − − +	.257	.736	45
+ − − −		.152	199

By increasing our information from three to four items we may learn that the second group, (+ + − −), belongs *after* the third group, (+ − − +), in the ordered scheme of response patterns. On the basis of three items only, we would have ordered these respondents in the reverse order. The reader is also invited to note that the finer scheme has a higher precision—fewer respondents get misclassified. *The effect of adding items* which satisfy the same conditions of reducibility *consists of improving the precision of a test.* This is due to the fact that attribution to latent classes is a probability procedure. We shall come back to this point at the end of this chapter.

From the examples in this section we have developed the following basic procedures in latent structure computations:

(a) The fitting of cross-product matrices for the purpose of evaluating the plausibility of a latent structure under consideration.

(b) The computation of the latent parameters—the latent structure scheme.

(c) The reproduction of ultimate frequencies as they derive from a given latent structure—the fitted response patterns.

(d) Exposition of the difference between fit and precision.

(e) The ordering of response patterns according to the probabilities that persons in each response pattern belong to the first latent class.

(f) Illustration of the role of adding to a system items which satisfy the original conditions of reducibility.

We shall see in the next section that very much the same notions apply also to other structures which are more complicated than the latent dichotomy case.

SECTION V

SOME SPECIAL LATENT STRUCTURES WITH MORE THAN TWO LATENT SUBCLASSES

1. A PARTIALLY DEGENERATE LATENT TRICHOTOMY

In Chapter 10, Section III, the solution of a "restricted quadratic case" was developed. One of the circumstances under which the case will often appear will be in surveys where a fairly large proportion of the respondents give responses of the "don't know," "undecided," or similar variety. Persons giving this type of response are most likely recruited from the middle range of the underlying attitude continuum; shifting about the "don't know's" or "undecided's" from the positive to the negative side in dichotomizing the responses to the questions will have the effect of changing the shape of the trace lines from convex to concave. Items which include the "don't know" response categories along with the positive responses to that question will have *convex* trace lines. The same items with the "don't knows" added to the negative response will have *concave* trace lines (see Section III, paragraph 3).

The example we shall use is four questions taken from a Research Branch questionnaire on job satisfaction. In the first two questions, the response category "undecided" has been classified with

the negative response; in the last two questions, the equivalent response categories have been classified with the positive reply.

1. Would you change to some other Army job if given a chance?
   Positive response: *No* (23.8%)
   Negative response: *Yes* and *Undecided* (76.2%)

3. Do you feel that everything possible has been done to place you in the Army job where you best fit?
   Positive response: *Yes* (32.5%)
   Negative response: *No* and *Undecided* (67.5%)

2. Do you usually feel that what you are doing in the Army is worthwhile or not?
   Positive response: *Worthwhile* and *Undecided* (72.9%)
   Negative response: *Not worthwhile* (27.1%)

4. How interested are you in your Army job?
   Positive response: *Very much interested* and *A little but not much* (83.3%)
   Negative response: *Not interested at all* (16.7%)

In line with the convention established previously, we have numbered the "concave" items with odd numbers and the "convex" items with even numbers.[5]

The empirical cross-product matrix of these four items is shown below in Table 8, with cross products standardized by division by n^2.

TABLE 8

EMPIRICAL CROSS-PRODUCT MATRIX OF JOB SATISFACTION ITEMS

	1	*3*	*2*	*4*
1	—	.0856	0375	.0267
3	.0856	—	.0411	.0423
2	.0375	.0411	—	.0667
4	.0267	.0423	.0667	—

The distinguishing characteristic of four-item cross-product matrices in the restricted quadratic case is the existence of one tetrad with zero or near-zero value and of another tetrad considerably greater or smaller than zero. The zero tetrad is always of the form:

[5] The notational conventions which we shall adhere to in discussing the restricted quadratic case are as follows:

	Concave items	*Convex items*
Item indices	odd	even
Subscript indices	i	j

(5.1) $$[i\bar{j}]\,[l\bar{k}] - [i\bar{k}]\,[\bar{j}l]$$

while the non-zero tetrad is always of the form:

(5.2) $$[il]\,[\bar{j}\bar{k}] - [i\bar{j}]\,[l\bar{k}]$$

In the present case, the tetrad $\left(\frac{[12]}{n^2}\frac{[34]}{n^2} - \frac{[14]}{n^2}\frac{[23]}{n^2}\right) = .00049$ which, while not quite zero, is quite small compared to the non-zero tetrad $\left(\frac{[13]}{n^2}\frac{[24]}{n^2} - \frac{[12]}{n^2}\frac{[34]}{n^2}\right) = .0041$ which is more than eight times as large.

Since items 1 and 3 are presumed to be "concave" items and items 2 and 4 to be "convex" items, the latent structure which generated the matrix has the form of Table 9:

TABLE 9

LATENT STRUCTURE OF RESTRICTED LATENT TRICHOTOMY

Latent class frequencies	*Latent marginals of items* 1	3	2	4
n_{I}	$a_1 + c_1$	$a_3 + c_3$	$a_2 + b_2$	$a_4 + b_4$
n_{II}	a_1	a_3	$a_2 + b_2$	$a_4 + b_4$
n_{III}	a_1	a_3	a_2	a_4

Four triplets of items may be formed from the matrix of four items. Two of the triplets will consist of two odd items and an even item, $\{132\}$ and $\{134\}$, and from these t_{I} may be computed. The other two triplets, $\{241\}$ and $\{243\}$, will consist of two even items and an odd one. From the latter pair of triplets, t_{III} can be computed. Incidentally, this rule obtains for matrices of any size in the restricted quadratic case.

From the triplet $\{132\}$ we can compute K_{I} as:

(5.3) $$K_{\mathrm{I}} = \frac{S_{1.23}}{p_1} - \frac{p_1}{S_{1.23}} + \frac{\eta_{1\,23}}{p_1 S_{1.23}} = \frac{S_{3.12}}{p_3} - \frac{p_3}{S_{3.12}} + \frac{\eta_{3\,12}}{p_3 S_{3.12}}$$

Note that it makes no difference whether the S- and η-values of item 1 or of item 3 are employed. The K is a characteristic of the triplet employed and does not depend (just as in the case of the latent dichotomy) on which concave item's S- and η-values are used.

Using item 1, we compute:

$$S_{1.23} = \sqrt{\frac{[13]\,[12]}{n^2\,[23]}} = \sqrt{\frac{(85{,}650)\,(37{,}498)}{1{,}000{,}000\,(41{,}075)}} = .2796$$

$$p_1 = \frac{238}{1{,}000} \doteq .2380$$

$$\eta_{1.23} = \frac{[23;1]}{[23]} = \frac{831}{41{,}075} = .0202$$

$$K_{\mathrm{I}} = \frac{.2796}{.2380} - \frac{.2380}{.2796} + \frac{.0202}{(.2380)\,(.2796)} = .6276$$

The triplet $\{134\}$ yields another value for K_{I}. In the case of perfect reducibility, K_{I} computed from $\{132\}$ will be identical with K_{I} computed from $\{134\}$. In the more usual case of imperfect data, the two values will differ somewhat.

Based on the triplet $\{134\}$ we find K_{I}:

$$S_{1\,34} = \sqrt{\frac{[13]\,[14]}{n^2\,[34]}} = \sqrt{\frac{(85{,}650)\,(26{,}746)}{1{,}000{,}000\,(42{,}275)}} = .2328$$

$$p_1 = \frac{238}{1000} = .2380$$

$$\eta_{1.34} = \frac{[34;1]}{[34]} = \frac{929}{42{,}275} = .0220$$

$$K_{\mathrm{I}} = \frac{.2328}{.2380} - \frac{.2380}{.2328} + \frac{.0220}{(.2380)\,(.2328)} = .3523$$

We take the arithmetic mean of the two estimates of K_{I}:

$$K_{\mathrm{I}} = \frac{.3523 + .6276}{2} = .4900$$

We can now compute $t_{\mathrm{I}} = \sqrt{\dfrac{n_{\mathrm{I}}}{n - n_{\mathrm{I}}}}$ by solving the equation

$$t_{\mathrm{I}}^2 + K_{\mathrm{I}} t_{\mathrm{I}} - 1 = 0 \tag{5.4}$$

$$t_{\mathrm{I}} = \frac{-K_{\mathrm{I}}}{2} + \sqrt{\frac{K_{\mathrm{I}}^2}{4} + 1} = .7846$$

The latent class frequency n_{I} can now be obtained from the relationship:

$$n_{\mathrm{I}} = \frac{nt_{\mathrm{I}}^2}{1 + t_{\mathrm{I}}^2} = 381.0 \tag{5.5}$$

K_{III} and the corresponding latent class frequency n_{III} are computed in the same way from the triplets $\{241\}$ and $\{243\}$. From the triplet $\{241\}$:

$$K_{\mathrm{III}} = \frac{S_{2\,14}}{p_2} - \frac{p_2}{S_{2.14}} + \frac{\eta_{2\,14}}{p_2 S_{2\,14}} = \frac{S_{4\,12}}{p_4} - \frac{p_4}{S_{4.12}} + \frac{\eta_{4.12}}{p_4 S_{4\,12}} \tag{5.6}$$

Computing K_{III} using item 2:

$$S_{2\,14} = \sqrt{\frac{[24]\,[12]}{n^2\,[14]}} = \sqrt{\frac{(66{,}743)\,(37{,}498)}{1{,}000{,}000\,(26{,}746)}} = .3059$$

$$p_2 = \frac{729}{1000} = .7290$$

$$\eta_{2.14} = \frac{[14;2]}{[14]} = \frac{7231}{26{,}746} = .2704$$

$$K_{\mathrm{III}} = \frac{.3059}{.7290} - \frac{.7290}{.3059} + \frac{.2704}{(.7290)\,(.3059)} = -.7512$$

From the triplet $\{243\}$:

$$K_{\mathrm{III}} = \frac{S_{2\,34}}{p_2} - \frac{p_2}{S_{2.34}} + \frac{\eta_{2.34}}{p_2 S_{2\,34}} = \frac{S_{4\,23}}{p_4} - \frac{p_4}{S_{4.23}} + \frac{\eta_{4.23}}{p_4 S_{4\,23}}$$

$$S_{2.34} = \sqrt{\frac{[24]\,[23]}{n^2\,[34]}} = \sqrt{\frac{(66{,}743)\,(41{,}075)}{1{,}000{,}000\,(42{,}275)}} = .2547$$

$$p_2 = \frac{729}{1000} = .7290$$

$$\eta_{2\,34} = \frac{[34;2]}{[34]} = \frac{13{,}103}{42{,}275} = .3099$$

$$K_{\mathrm{III}} = \frac{.2547}{.7290} - \frac{.7290}{.2547} + \frac{.3099}{(.7290)\,(.2547)} = -.8438$$

Taking the arithmetic average of these two estimates of K_{III}:

$$K_{\mathrm{III}} = \frac{-.7512 - .8438}{2} = -.7975$$

Solving for $t_{\mathrm{III}} = \sqrt{\dfrac{n - n_{\mathrm{III}}}{n_{\mathrm{III}}}}$ in the formula

(5.7) $$t_{\mathrm{III}}^2 + K_{\mathrm{III}}t_{\mathrm{III}} - 1 = 0$$

$$t_{\mathrm{III}} = \frac{-K_{\mathrm{III}}}{2} + \sqrt{\frac{K_{\mathrm{III}}^2}{4} + 1} = 1.4753$$

We get

(5.8) $$n_{\mathrm{III}} = \frac{n}{1 + t_{\mathrm{III}}^2} = 314.8$$

The remaining latent class frequency n_{II} may be found directly, once n_{I} and n_{III} are known, since

$$n_{\mathrm{II}} = n - n_{\mathrm{I}} - n_{\mathrm{III}} = 1{,}000 - 381.0 - 314.8 = 304.2$$

Just as we adjust the values of K_{I} and K_{III} by pooling the two estimates of each K obtained in each pair of triplets, so also we adjust the S-values in computing the item latent marginals. For each item there are two S-values, one from each triplet relevant to that item. For item 1, e.g., we get $S_{1\,23}$ based on the triplet $\{123\}$ and $S_{1\,34}$ based on the triplet $\{134\}$. For lack of any rigid sampling theory, we take the geometric mean of the various S-values and compute an adjusted S'.

(5.9)
$$S_1' = \frac{1}{n}\sqrt[4]{\frac{[13]^2\,[12]\,[14]}{[23]\,[34]}} = .2551$$
$$S_3' = \frac{1}{n}\sqrt[4]{\frac{[13]^2\,[23]\,[34]}{[12]\,[14]}} = .3357$$
$$S_2' = \frac{1}{n}\sqrt[4]{\frac{[24]^2\,[12]\,[23]}{[14]\,[34]}} = .2791$$
$$S_4' = \frac{1}{n}\sqrt[4]{\frac{[24]^2\,[14]\,[34]}{[12]\,[23]}} = .2391$$

The a-values for each item are given from the relationships

(5.10)
$$a_i = p_i - S_i' t_{\mathrm{I}}$$
$$a_j = p_j - S_j' t_{\mathrm{III}}$$

and the c- and b-values from the formulae

(5.11)
$$c_i = \frac{S_i'}{\sqrt{\nu_{\mathrm{I}}\,(1 - \nu_{\mathrm{I}})}} \qquad \text{where} \quad \nu_{\mathrm{I}} = \frac{n_{\mathrm{I}}}{n}$$
$$b_j = \frac{S_j'}{\sqrt{\nu_{\mathrm{III}}\,(1 - \nu_{\mathrm{III}})}} \qquad \text{where} \quad \nu_{\mathrm{III}} = \frac{n_{\mathrm{III}}}{n}$$

The a-values in the present case are easily computed using the formulae (5.10):

$$
\begin{aligned}
a_1 &= p_1 - S_1' t_{\mathrm{I}} &&= .2380 - (.2551)\,(.7846) &&= .0378\\
a_3 &= p_3 - S_3' t_{\mathrm{I}} &&= .3250 - (.3357)\,(.7846) &&= .0616\\
a_2 &= p_2 - S_2' t_{\mathrm{III}} &&= .7290 - (.2791)\,(1.4753) &&= .3172\\
a_4 &= p_4 - S_4' t_{\mathrm{III}} &&= .8330 - (.2391)\,(1.4753) &&= .4803
\end{aligned}
$$

Finally, the b- and c-values, from (5.11):

$$c_1 = \frac{S_1'}{\sqrt{\nu_{\mathrm{I}}\,(1-\nu_{\mathrm{I}})}} = \frac{.2551}{\sqrt{(.3810)\,(.6190)}} = .5253$$

$$c_3 = \frac{S_3'}{\sqrt{\nu_{\mathrm{I}}\,(1-\nu_{\mathrm{I}})}} = \frac{.3357}{\sqrt{(.3810)\,(.6190)}} = .6913$$

$$b_2 = \frac{S_2'}{\sqrt{\nu_{\mathrm{III}}\,(1-\nu_{\mathrm{III}})}} = \frac{.2791}{\sqrt{(.3148)\,(.6852)}} = .6011$$

$$b_4 = \frac{S_4'}{\sqrt{\nu_{\mathrm{III}}\,(1-\nu_{\mathrm{III}})}} = \frac{.2391}{\sqrt{(.3148)\,(.6852)}} = .5149$$

The latent structure for the present case is found by combining the a-, b-, and c-values according to the hypothetical latent structure and is shown below in Table 10.

TABLE 10

COMPUTED LATENT STRUCTURE FROM JOB SATISFACTION ITEMS

Latent class frequencies	*Latent marginals for items* 1	3	2	4
$n_{\mathrm{I}} = 381.0$	.563	.753	.918	.995
$n_{\mathrm{II}} = 304.2$	.038	.062	.918	.995
$n_{\mathrm{III}} = 314.8$	.038	.062	.317	.480

2. CLASSIFYING RESPONDENTS ACCORDING TO A LATENT TRICHOTOMY

In order to use the latent structure for ordering response patterns, we have now to compute from Table 10 the ultimate frequencies corresponding to the mathematically perfect case. In doing so we will at the same time learn how the people who make these responses recruit themselves from the latent classes. Since we have three latent subclasses, persons giving each response pattern will be divided into three homogeneous classes and the number of persons

giving a certain response pattern can be computed by multiplying the appropriate pattern of positive latent marginals and complements of positive latent marginals in each subclass.

The number of persons giving the response pattern $(- - + +)$ recruited from latent class I is:

$$n_{\mathrm{I}}\,(1 - p_{\mathrm{I}1})\,(1 - p_{\mathrm{I}3})\,p_{\mathrm{I}2}p_{\mathrm{I}4} = (381.0)\,(.437)\,(.247)\,(.918)\,(.995) = 37.59$$

The number of persons giving this response pattern recruited from latent class II is:

$$n_{\mathrm{II}}\,(1 - p_{\mathrm{II}1})\,(1 - p_{\mathrm{II}3})\,p_{\mathrm{II}2}p_{\mathrm{II}4} = (304.2)\,(.962)\,(.938)\,(.918)\,(.995) = 251.0$$

Finally, the number of persons recruited from latent class III is given by:

$$n_{\mathrm{III}}\,(1 - p_{\mathrm{III}1})\,(1 - p_{\mathrm{III}3})\,p_{\mathrm{III}2}p_{\mathrm{III}4} = (314.8)\,(.962)\,(.938)\,(.317)\,(.480) = 43.30$$

By adding these three components we get 331.9 as a fitted frequency. The other figures in column VI of Table 11 were obtained by similar procedures. The reader will notice that a simple rule leads from Table 10 to Table 11. We first multiply n_{I} by the four probabilities or their complements in the first row of Table 10. The same operation is repeated in the second and third row using n_{II} and n_{III}. Then an addition over the three rows is performed. Whether for an item p_{I_j} or $(1 - p_{\mathrm{I}_j})$ is used depends on whether, in the response pattern in question, the answer to this item is positive or negative.

In columns II, III, and IV of Table 11 we have not entered the absolute numbers obtained. Rather, the number in each class is expressed as a per cent of all the respondents giving a certain response pattern (column VI). Row by row, columns II, III, and IV add to 100 per cent in Table 11. These are the a posteriori probabilities discussed in paragraph 5 of the preceding section. They indicate *the probability that a respondent giving a particular response pattern comes from each one of the three latent subclasses.*

As in the dichotomous case, we can classify each response pattern according to its modal probability. Reading from top to bottom in Table 11, the people who give the first seven response patterns will most likely come from latent class I, the next from class II, and the last eight from class III. Within these groups, we can perform a

further ordering, but it will remain somewhat arbitrary. The problem is exemplified by a comparison of patterns (− − − +) and (− + + −). According to column IV, the former has a lower probability for class III than the latter. However (− + + −) might be called "bimodal." It has a sizable recruitment from class I, which (− − − +) does not have. To order these two patterns requires some convention such as assigning numerical values to the

TABLE 11

COMPLETE ANALYSIS OF RESTRICTED LATENT TRICHOTOMY ON JOB SATISFACTION

I	*II*	*III*	*IV*	*V*	*VI*	*VII*
Response pattern	*Per cent in latent class* *I*	*II*	*III*		*Fitted total*	*Actual total*
+ + + +	99.5%	0.4%	0.1%	100%	148.4	147
+ + − +	97.8	0.4	1.8	100%	13.4	11
− + + +	85.6	12.3	2.1	100%	133.9	128
+ + + −	85.2	0.4	14.4	100%	0.8	1
+ − + +	80.7	16.4	2.9	100%	60.0	58
− + − +	57.3	8.3	34.4	100%	17.8	27
+ − − +	48.7	9.9	41.4	100%	8.9	9
− − + +	11.3	75.6	13.1	100%	331.9	341
− − − +	2.8	18.8	78.4	100%	118.9	112
+ + − −	19.6	0.1	80.3	100%	0.3	4
− + + −	14.9	2.1	83.0	100%	3.7	2
+ − + −	11.0	2.2	86.8	100%	2.1	5
− − + −	0.4	2.5	97.1	100%	48.2	47
− + − −	0.7	0.1	99.2	100%	6.7	5
+ − − −	0.5	0.1	99.4	100%	4.0	3
− − − −	0.0	0.1	99.9	100%	101.0	100
Totals in each class	381.0	304.2	314.8		1,000.0	1,000

latent classes and then using the average of the latent marginals for each response pattern as an index. We shall come back to this point twice, in paragraphs 2 and 3 in the last section.

Here, as in the dichotomous case, we can develop a *measure of precision:* the proportion of people who are properly classified by this structure. The original data, e.g., shows that of the 133.9 which in the perfect case give response pattern (− + + +), 114.5 come from class I, 16.5 from class II, and 2.9 from class III. But we cannot tell which individuals these are. All we know is their identical response pattern (− + + +). The best solution is to

put all 133.9 in class I. But only 114.5 belong there. Repeating this reasoning for each line, we obtain a precision of 84.7 per cent. This means that 15.3 per cent are inevitably misclassified.

The reader can check for himself that the precision would be considerably less if only three items had been used. Inversely, the adding of further items, satisfying the conditions of reducibility, would increase the precision of the test. Experiments are under way to study how this precision increases. It also would be desirable to study further the mathematical relationship between the number of items and degree of precision. The formulae are easily derived but lead to rather involved combinatorial forms.

Finally, the reader is asked to compare columns VI and VII of Table 11. The "fit" of the structure to the empirical data can be gauged from these two columns of figures. In the present example, the fit is rather close. There is nothing in the interpretation of closeness of a trichotomous fit which goes beyond that discussed in the dichotomous case.

3. THE LATENT DISTANCE CASE

The more latent classes there are, the more will it be necessary to develop general ideas as to how a latent structure can be used to order manifest response patterns. The next example will therefore deal with the "latent distance" case, which of all the cases dealt with at the end of the previous chapter has the largest number of latent subclasses. For a discussion we turn to a concrete example. The following four items were taken from a neurotic inventory:

1. Have you ever been bothered by pressure or pains in the head?
   Positive answer: *Yes, often* or *Yes, sometimes* or *No answer* 13.8%
2. Have you ever been bothered by shortness of breath when you were not exercising or working hard?
   Positive answer: *Yes, often* or *Yes, sometimes* or *No answer* 30.7
3. Do your hands ever tremble enough to bother you?
   Positive answer: *Yes, often* or *Yes, sometimes* or *No answer* 43.1
4. Do you often have trouble in getting to sleep or staying asleep?
   Positive answer: *Very often* or *No answer* 57.1

The assumption was made that a latent distance structure would fit the material. Therefore the nine parameters of Table 12 had to be computed:

TABLE 12

LATENT STRUCTURE OF LATENT DISTANCE SCALE

Latent class frequencies	*Latent marginals of items* 1	2	3	4
n_{I}	$1 - d_1$	$1 - d_2$	$1 - d_3$	$1 - d_4$
n_{II}	d_1	$1 - d_2$	$1 - d_3$	$1 - d_4$
n_{III}	d_1	d_2	$1 - d_3$	$1 - d_4$
n_{IV}	d_1	d_2	d_3	$1 - d_4$
n_{V}	d_1	d_2	d_3	d_4

It is immediately obvious that here the items can be arranged in terms of their manifest marginals. For the value $n_i/n = x_i(1 - d_i) + (1 - x_i)d_i$ obviously increases in the order in which the items are arranged in Table 12. (The reader will remember that x_i indicates the "breaking point" of each item—see Section III, paragraph 5.)

In order to compute the latent parameters, we can discard the equations for n_{1234} and for the joint frequencies of three items. This leaves eleven equations for nine parameters and therefore two further conditions have to be satisfied. Here the propositions of Section II, paragraph 6, Chapter 10, are of no help because $\lambda > m$ $(5 > 4)$. Special computations have to be made to find these two conditions. One is again a tetrad condition:

$$[12,34] = [13]\,[24] - [14]\,[23] = 0$$

which means that the tetrads *on one side* of the principal diagonal vanish (but *not* those which cross the principal diagonal). From this type of condition we again derive adjusted S-values as geometric means of the values

$$S_i = \sqrt{\frac{[ia]\,[ib]}{n^2\,[ab]}} \tag{5.12}$$

But here a and b, the two auxiliary items, have to satisfy the condition that $a > i > b$ (where items are ordered by their manifest marginals).

For the present structure, therefore, S-values are defined only for the "medial" items (in our example the medial items are item 2 and item 3) excluding the first and the last. Once adjusted S-values have been computed, it is easy to compute all the d-values of the medial items from the equations

$$(p_i - d_i)(q_i - d_i) = S_i^2 \tag{5.13}$$

In our present example we find

$[12] = 53{,}634$	$[13] = 45{,}522$	$[14] = 31{,}202$
$[23] = 71{,}683$	$[24] \doteq 68{,}703$	$[34] = 93{,}899$

Therefore

$$S_2^2 = \frac{1}{n^2}\sqrt{\frac{[12]\,[23]\,[12]\,[24]}{[13]\,[14]}} = .099870$$

$$S_3^2 = \frac{1}{n^2}\sqrt{\frac{[13]\,[34]\,[23]\,[34]}{[14]\,[24]}} = .115851$$

And finally

$$d_2 = .129705 \qquad\qquad d_3 = .152708$$

as the smaller roots of the two quadratic equations

$$d_2^2 - d_2 + (p_2q_2 - S_2^2) = 0$$

$$d_3^2 - d_3 + (p_3q_3 - S_3^2) = 0$$

which are solved, according to the quadratic rule,

$$d_2 = \frac{1}{2} - \sqrt{\frac{1}{4} - (p_2q_2 - S_2^2)} = \frac{1}{2} - \sqrt{\frac{1}{4} - (.307 \times .693 - .099870)}$$

$$d_3 = \frac{1}{2} - \sqrt{\frac{1}{4} - (p_3q_3 - S_3^2)} = \frac{1}{2} - \sqrt{\frac{1}{4} - (.431 \times .569 - .115851)}$$

This leaves d_1 and d_m (in our case d_4) to be computed, and here the second condition mentioned above comes in. This second condition is complicated to put in manifest terms, but it derives from the fact that

$$p_1 - d_1 = \frac{[12]}{n^2(q_2 - d_2)} = \frac{[13]}{n^2(q_3 - d_3)}$$

or the equivalent form

$$q_4 - d_4 = \frac{[24]}{n^2(p_2 - d_2)} = \frac{[34]}{n^2(p_3 - d_3)}$$

Again, if we have to work with imperfect data we will put

$$d_1 = p_1 - \frac{1}{n^2}\sqrt{\frac{[12]\,[13]}{(q_2 - d_2)\,(q_3 - d_3)}}$$

$$d_4 = q_4 - \frac{1}{n^2}\sqrt{\frac{[24]\,[34]}{(p_2 - d_2)\,(p_3 - d_3)}}$$

In our example we now know from the previous step that

$$q_2 - d_2 = .563295, \qquad p_2 - d_2 = .177295$$
$$q_3 - d_3 = .416292, \qquad p_3 - d_3 = .278292$$

Therefore

$$d_1 = .138 - \frac{1}{1000^2}\sqrt{\frac{53{,}634 \times 45{,}522}{.563295 \times .416292}} = .0359616$$

$$d_4 = .429 - \frac{1}{1000^2}\sqrt{\frac{68{,}703 \times 93{,}899}{.177295 \times .278292}} = .0674072$$

Once all the d-values have been found we get immediately the latent class frequencies from formula (3.19) in Section III.

$$x_1 = \frac{p_1 - d_1}{1 - 2d_1} = \frac{.138 - .0359616}{1 - 2(.0359616)} = .109946$$

$$x_2 = \frac{p_2 - d_2}{1 - 2d_2} = \frac{.307 - .129705}{1 - 2(.129705)} = .239397$$

$$x_3 = \frac{p_3 - d_3}{1 - 2d_3} = \frac{.431 - .152708}{1 - 2(.152708)} = .400660$$

$$x_4 = \frac{p_4 - d_4}{1 - 2d_4} = \frac{.571 - .0674072}{1 - 2(.0674072)} = .582063$$

$$\begin{aligned}
n_{\mathrm{I}} &= nx_1 &&= 109.95\\
n_{\mathrm{II}} &= nx_2 - nx_1 &&= 129.45\\
n_{\mathrm{III}} &= nx_3 - nx_2 &&= 161.26\\
n_{\mathrm{IV}} &= nx_4 - nx_3 &&= 181.40\\
n_{\mathrm{V}} &= n \quad - nx_4 &&= 417.94
\end{aligned}$$

We are now in a position to judge the fit of this structure and at the same time to see how each response pattern recruits itself from the different latent subclasses. In Table 13, the ultimate fitted frequencies are computed in the usual fashion from the latent mar-

ginals and latent class frequencies which have now been derived for Table 12: i.e., n_{1234} $(+ + + +)$ recruits $n_{\text{IV}}\, d_1\, d_2\, d_3\, (1 - d_4)$ cases from subclass IV, and $n_{\text{V}}\, d_1\, d_2\, d_3\, d_4$ cases from subclass V. Of course, the computation of the fitted frequencies for a response pattern with a negative response for an item requires the use of the complement of the latent marginal derived for that item in Table 12. Table 13 presents the results in percentages for convenience.

TABLE 13

COMPLETE ANALYSIS OF LATENT DISTANCE SCALE ON NEUROTIC INVENTORY

Response pattern *1 2 3 4*	*Per cent of each pattern in latent class:* n_{I}	n_{II}	n_{III}	n_{IV}	n_{V}		*Fitted total*	*Actual total*
+ + + +	94.9%	4.2%	0.8%	0.1%	0.0%	100%	76.8	75
+ + − +	90.0	3.9	0.7	4.6	0.8	100%	14.6	10
+ + + −	90.4	4.0	0.7	0.1	4.8	100%	5.8	8
+ − + +	66.8	2.9	24.5	5.0	0.8	100%	16.3	14
− + + +	2.5	79.3	14.7	3.0	0.5	100%	108 3	110
− − + +	0.3	8.8	73.5	14.9	2.5	100%	145.4	141
− + − +	1.3	38 9	7.2	45.1	7.5	100%	39.7	49
+ − − +	24.5	1.1	9.0	56.1	9.3	100%	8.0	11
− − − +	0.0	1.4	11.9	74.3	12.4	100%	161.9	161
+ + − −	36.7	1.6	0.3	1.9	59.5	100%	2.6	3
− + + −	1.3	40.7	7.6	1.5	48.9	100%	15.2	11
+ − + −	25.9	1.2	9.5	1.9	61.5	100%	3.0	8
− − + −	0.1	1.5	12.8	2.6	83.0	100%	60.2	64
− + − −	0.1	2.5	0.5	2.9	94.0	100%	44.0	41
+ − − −	1.3	0.0	0.5	3.0	95.2	100%	10.9	9
− − − −	0.0	0.1	0.5	3.0	96.4	100%	287.3	285
Total in each class	109 9	129.5	161.3	181.4	417.9		1,000.0	1,000

There are certain response patterns like $(+ + + +)$ where most cases come from one latent subclass. Others like $(- - - +)$ show a latent distribution with much greater dispersion but still with an obvious peak in one latent subclass. But some patterns show a bimodal latent distribution; responses of this kind seem to be ambiguous. They come from people who can have quite different positions on the latent "scale" which underlies the manifest data, e.g., the 49 cases who answer $(- + - +)$.

It is of course possible to rank response patterns by some kind of average latent position: the mode or the median would not require

further assumptions; the arithmetic mean would require that we arbitrarily assign numerical values, say 1 to 5, to the latent subclasses. But one can also state that certain response patterns with a bimodal latent distribution should not be included in the scale. People who score this way cannot be properly ranked. Finally, it is possible to say that a variety of indices could be used to characterize each response pattern. In addition to the average, we would use and interpret, e.g., the standard deviation of the latent distribution. This last procedure would correspond to suggestions recently made by Clyde Coombs.[6]

4. THE NOTION OF LATENT LINEARITY

In the last two examples we were concerned with the problem of how a set of latent classes, arranged in a certain order, can be used to give order to the manifest response patterns of a dichotomous system. In this way, the first connection between latent structure analysis and the problem of scaling was established; for the purpose of any scaling model is to provide a device by which manifest response patterns can be ordered in some rational way. The discussion, however, was predicated on the assumption that the latent classes were themselves ordered. But nothing in the algebra of latent structure analysis provides for an ordering of the latent classes. Therefore, an additional *criterion of latent linearity* has to be introduced. The following definition will be used:

Assume a set of latent classes, n_{I}, n_{II}, $\cdots$; assume that for a given item j, the classes can be so arranged that, up to a certain class x_j, the difference of two previous latent marginals, $p_{(a+1)j} - p_{(a)j} \geqq 0$; for all subsequent classes, $p_{(b)j} - p_{(b-1)j} \leqq 0$ (where $a < x_j < b$); call the class x_j the "maximum class" for item j; if then each item has a maximum class for the same order of the latent classes, we will call this a linear order of latent classes. The structure will be said to have latent linearity. A schematic example of such an arrangement is presented in Table 14.

The latent structure of the restricted trichotomy, investigated in paragraph 2 of this section, and the latent distance case, discussed in paragraph 3, are examples of structures with latent linearity.

Before giving an example of nonlinearity the notion of latent linearity should be related to the concept of trace lines. In Section I no limitation was put on the kind of trace lines which could exist.

[6] Clyde H. Coombs, "Some Hypotheses For the Analyses of Qualitative Variables," *Psychological Review*, Vol. 55, No. 3 (May 1948), pp. 167–174.

They could have several maxima, oscillate back and forth, or do anything else which could be expressed by polynomials of a high degree. It makes psychological sense, however, to single out such trace lines which have at most one maximum over the populated segment of the x-continuum. We can call these trace lines unimodal; they will exclude U-shaped curves and will be either monotonic ascending, monotonic descending or bell-shaped. Then the following characterization holds true: *the equivalent latent class structures of unimodal trace lines have latent linearity.*

A large variety of structures fall into this category. Consider, e.g., the possible combinations of the latent distance case and the

TABLE 14

SCHEMATIC EXAMPLE OF LATENT LINEARITY

Latent class frequencies	*Latent marginals for items* 1	2	3	4
$n_{I} = 100$	*.90*	.50	.40	.20
$n_{II} = 300$	.70	*.80*	.60	.30
$n_{III} = 400$	.40	.50	*.90*	.40
$n_{IV} = 200$	.20	.30	.40	*.80*

special latent trichotomy which were just investigated. By looking at the structures of paragraphs 2 and 3 in this section, we see that the latent trichotomy could really be looked upon as a two-step latent distance case, with two questions for each step. This suggests that one of the structures to be investigated in the future should be a general latent distance scale, with several items on each step, which would then form a series of ordered latent dichotomies. On the other hand, a general scheme like that in Table 14 could be approximated by an adaptation of the scheme in paragraph 3. This would give a manageable way to analyze the Thurstone attitude scale.

Now we shall give a concrete example of a case where latent linearity does not prevail.

5. THE LATENT TURNOVER

In paragraph 4 of Section III the case of two latent dichotomies was discussed. The mathematics developed there applies quite easily, for instance, to repeated interviews. The questions asked in the first interview may form one latent dichotomy; the questions asked in the second interview a second latent dichotomy. The

replies to the individual questions might show a considerable turnover from the first to the second interview. But the latent attribute which is common to all the questions might show a great consistency from one interview to the next.

We give here an example from a repeated interview study done by the Research Branch on attitudes toward officers. Three questions were asked:[7]

	INTERVIEW 1st	INTERVIEW 2nd
1. How many of your officers are fair-minded enough to treat anybody in the outfit alike? Positive answer: *All* or *Most*	79%	59%
3. How many of the officers you now serve under are the kind you would want to serve under in combat? Positive answer: *All* or *Most*	65	36
5. How many of your company officers do you think would be willing to go through anything they ask their men to go through? Positive answer: *All* or *Most*	65	49

(The results of these items on the second interview will be referred to as items 2, 4, and 6.)

The first interview gave the following manifest data:

$$n = 324 \qquad n_{13} = 197$$
$$n_1 = 257 \qquad n_{15} = 194$$
$$n_3 = 211 \qquad n_{35} = 172$$
$$n_5 = 209 \qquad n_{135} = 165$$

Applying the method of solution to two latent classes as given previously, we find:

$$p_1 = .793210 \qquad p_3 = .651235 \qquad p_5 = .645062$$
$$S_1 = .268155 \qquad S_3 = .341067 \qquad S_5 = .324797$$
$$\mu_1 = .360048 \qquad \mu_3 = .101827 \qquad \mu_5 = .116342$$

and by the formula

$$K = \frac{S_i}{p_i} - \frac{p_i}{S_i} + \frac{\mu_i}{S_i p_i}$$

[7] Note that these questions are numbered with odd indices on the first interview and even indices on the second interview. This convention will be employed throughout this discussion.

we obtain

$$K = -.927238$$

and

$$t = -\frac{K}{2} + \sqrt{\frac{K^2}{4} + 1} = 1.56586$$

From t we can compute the entire latent structure:

Latent class frequencies		*Latent marginals* (1)	(3)	(5)
n_A	230.14	.96446	.86905	.85249
$n - n_A$	93.86	.37332	.11717	.13648

The second interview gives for manifest data:

$$\begin{aligned} n &= 324 & n_{24} &= 104 \\ n_2 &= 191 & n_{26} &= 133 \\ n_4 &= 118 & n_{46} &= 105 \\ n_6 &= 160 & n_{246} &= 97 \end{aligned}$$

The auxiliary parameters are:

$$\begin{aligned} p_2 &= .589506 & p_4 &= .364198 & p_6 &= .493827 \\ S_2 &= .296616 & S_4 &= .358344 & S_6 &= .402471 \\ \mu_2 &= .310106 & \mu_4 &= .0419726 & \mu_6 &= .139362 \end{aligned}$$

$$K = .289205$$
$$t = .86580$$

Latent class frequencies		*Latent marginals* (2)	(4)	(6)
n_B	138.82	.93210	.77809	.95868
$n - n_B$	185.18	.33270	.05394	.14537

We need now the four latent class frequencies to complete the latent structure. These form a fourfold table:

		Latent attribute A (attitude on 1st interview) +	−	
Latent attribute B (attitude on 2nd interview)	+	n_{I}	n_{III}	$n_{\text{I}} + n_{\text{III}} = n_B = 138.8$
	−	n_{II}	n_{IV}	$n_{\text{II}} + n_{\text{IV}} = 185.2$
		$n_{\text{I}} + n_{\text{II}} = n_A = 230.1$	$n_{\text{III}} + n_{\text{IV}} = 93.9$	$n = 324$

The marginals of this fourfold table are the latent class frequencies computed for each interview separately, as was done above. To fill in the values n_{I}, n_{II}, n_{III} and n_{IV}, the numbers who have all the four possible combinations of positions on the two interviews taken together, we need only compute the cross product $[AB]$ of this latent fourfold table.

To do this we turn to the cross-product matrix, Table 15, formed with all six items. The cross products have been standardized by division by n^2.

TABLE 15

STANDARDIZED CROSS-PRODUCT MATRIX OF REPEATED INTERVIEW STUDY: $[ij]/n^2$

	1	*3*	*5*	*2*	*4*	*6*
1	——	**0915**	**.0871**	*.0694*	*0506*	*0496*
3	.0915	——	**.1108**	*.0451*	*.0591*	*.0457*
5	0871	.1108	——	*0703*	*0614*	*0673*
2	.0694	.0451	.0703	——	**.1063**	**.1194**
4	.0506	0591	0614	1063	——	**1443**
6	.0496	0457	.0673	1194	.1443	——

Attention is directed to the nine cross products (in italics) in the right upper corner. Each is formed between an item from the first interview and one from the second. It is known from Section III, paragraph 4, that the cross product $[AB]$ between the two latent dichotomies can in the mathematically perfect case be computed from any of the nine combinations:

$$[AB] = \frac{1}{S_i S_{\bar{k}}} \frac{[i\bar{k}]}{n^2} \sqrt{n_A (n - n_A) n_B (n - n_B)} \tag{5.14}$$

where as usual the barred index pertains to an even item (second interview); the unbarred to an odd item (first interview).

With fallible data the nine values of (5.14) will give slightly different results and again a geometric mean will be used to get an adjusted $[AB]$. If we have odd and even items, then the following formula can easily be verified:

(5.15)

$$\log [AB] = \frac{1}{h\bar{g}} \sum \log [i\bar{k}] - \frac{1}{h(h-1)} \sum \log [il] - \frac{1}{\bar{g}(\bar{g}-1)} \sum \log [\bar{k}\bar{j}] + \log \frac{n_A (n - n_A) n_B (n - n_B)}{2}$$

h = number of odd items, $\bar{g}$ = number of even items

The first sum on the right is taken over all the italicized cross products of Table 15 (after conversion into logarithms). The second and the third sums are taken in the same way over the two blocks of boldface figures which cover the odd and the even items respectively.

In the present case, $h = \bar{g} = 3$. Therefore, the adjusted $[AB]$ becomes

$$[AB] = \sqrt{n_A (n - n_A) n_B (n - n_B)} \frac{\sqrt[9]{[12]\,[14]\,[16]\,[23]\,[34]\,[36]\,[25]\,[45]\,[56]}}{\sqrt[6]{[13]\,[15]\,[35]\,[24]\,[26]\,[46]}}$$

At this point a correction has to be made due to the fact that these six items are in fact the same three questions asked on two different occasions. This seems likely to produce a special relationship between the respondents' answers to the same question when asked on the two occasions, in excess of the relation due to general attributes. We shall correct for this by removing the three cross products which arise from the repeated use of the same item: [12], [34], and [56]; for the numerator of the formula we may take the sixth root of the remaining six cross products.

Applying this for our example (through use of logarithms), we obtain the value of the cross product of the latent fourfold table:

$$[AB] = 11{,}640$$

and therefore

$$n_{\mathrm{I}} = \frac{n_A n_B + [AB]}{n} = \frac{230.1 \times 138.8 + 11{,}640}{324} = 134.5$$

We can now fill in the four latent class frequencies:

		First interview +	*First interview* −	
Second interview	+	n_{I} 134.5 (41.5%)	n_{III} 4.3 (1.3%)	138.8
	−	n_{II} 95.6 (29.5%)	n_{IV} 89.6 (27.7%)	185.2
		230.1	93.9	324 = 100%

This result suggests that only about four cases have changed their basic attitude from negative to positive, while about ninety-six cases have gone from a favorable to an unfavorable attitude, the other two groups remaining constant in their attitude.

By contrast, the turnover on the three individual items was:

		Item 1 *First interview* +	−		*Item 3* *First interview* +	−		*Item 5* *First interview* +	−
Second interview	+	53.8%	5.2%	+	29.6%	6.8%	+	38.6%	10.8%
	−	25.6%	15.4%	−	35.5%	28.1%	−	25.9%	24.7%

On individual questions, the number who went counter to the trend from negative to positive ranges from 5 to 10 per cent. In terms of the latent attribute, this "backwash" amounted to only 1.3 per cent. The two groups who did not change are both larger on the latent attribute turnover than the average of the individual questions. The latent attribute appears to be generally less subject to chance fluctuations one way or the other.

Finally, we can record the whole latent structure in Table 16 and return to the discussion of latent linearity.

TABLE 16

LATENT STRUCTURE OF REPEATED INTERVIEW STUDY

Latent class frequencies	*Latent marginals for items* 1	3	5	2	4	6
n_{I} = 134 5	964	.869	852	.932	.778	.959
n_{II} = 95.6	.964	.869	.852	333	.054	.145
n_{III} = 4.3	.373	117	.136	.932	.778	959
n_{IV} = 89.6	.373	.117	.136	.333	.054	.145

6. CASES WITH MORE THAN ONE UNDERLYING CONTINUUM

Obviously, the structure of Table 16 does not permit an arrangement of the latent subclasses as requested by the definition of latent linearity in paragraph 4 above. The even items have a bimodal sequence of latent marginals. If we were to rearrange the order of the latent subclasses, so as to assure a monotonic sequence for the even items, the odd items would have a bimodal sequence of the latent marginals. The fact is easily explained by the nature of the example: it is the result of having assumed *two* latent dichotomies. But the finding can be reversed, and then it suggests a generalization:

Suppose we had analyzed the original data blindly, merely guided by their manifest rank structure, and had obtained a latent structure scheme like Table 16. Then, two interpretations would have been open to us. We could have said (a) that this is a four-class structure equivalent to one underlying continuum; or (b) that it represents two latent dichotomies. In the first case, we would have to add that the trace lines are not unimodal. Which of these two interpretations is preferable could be decided only in the light of the subject matter under investigation. However a review of current test practices suggests that the second interpretation would be preferred in most concrete problems. As a matter of fact, notions such as "unidimensionality" or "homogeneity" of tests seem well covered by our concept of latent linearity. But the fact remains that it implies an additional convention and cannot be derived from a basic mathematical formulation.

To round out the picture, we shall reformulate the matter in mathematical terms, restricting ourselves for simplicity's sake to two variables. Suppose that there were two underlying continua, x and z. People then could be located in an (x,z) plane. A dichotomous question would have a trace surface $y = f(x,z)$, indicating the probability of a positive answer at each point of the underlying continuum plane. The position of all respondents would be described by a distribution surface $y = \phi(x,z)$. The accounting equations, linking this model with the manifest joint occurrences and response patterns, would be:

$$(5.16) \qquad p_{ij\ldots\ldots} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_i(x,z)\, f_j(x,z) \cdots \phi(x,z)\,dx\,dz$$

As in Section I, additional assumptions will have to be made to make these equations more specific and accessible to empirical solution. Suppose that the trace surfaces in a specific case are quadratic; then it would be possible to partition the (x,z) plane into nine rectangular sectors and to assume that in each sector the population is concentrated at one point. Thus we would get nine segmental probabilities so that:

$$(5.17) \qquad p_{ij\cdots\cdot} = n_{\mathrm{I}} p_{\mathrm{I}i} p_{\mathrm{I}j} \cdots + n_{\mathrm{II}} p_{\mathrm{II}i} p_{\mathrm{II}j} \cdots + \cdots$$

These equations have a solution which can be interpreted as follows: the latent class frequencies $n_{\mathrm{I}}, n_{\mathrm{II}}, \cdots, n_{\mathrm{IX}}$ are the number of people concentrated at one point in each partition of the (x,z) plane. Within each partition the responses to the test items are independent of each other. Each test item has in each sector of the plane one latent probability. The response patterns of the whole sample are related to these latent parameters by accounting equations of the type (5.17).

But obviously the solutions of (5.17) can also be viewed as belonging in the frame of reference of the discussion in Section I of Chapter 10. They might as well refer to *one* underlying continuum divided into nine segments. Which of the two interpretations mentioned should be used cannot be decided on mathematical grounds. But what is important is this: if it is desirable to assume two or more underlying continua, then the algebra developed in Section II of Chapter 10 is adequate for dealing with this more complicated structure also.

SECTION VI

LATENT STRUCTURE ANALYSIS AND TEST CONSTRUCTION

In the last two sections we have chiefly developed computing procedures for the derivation of latent parameters from manifest data. As we proceeded, a number of incidental observations were made, especially on how the latent parameters could be used to bring order into the manifest response patterns. It now seems advisable to discuss in a more general way the relation between the present approach and a few topics from the field of test construction.

From the beginning it should be clear that latent structure theory does not propose the use of one specific test procedure in preference to others. Its purpose is merely to bring out clearly and precisely

what seems to be common to a large variety of techniques. The discussion below is concerned not with deciding which of several procedures is best, but rather with how they are related to one another.

1. THE LATENT STRUCTURE OF THE SCALOGRAM PATTERN

A considerable portion of the present volume is devoted to the specific types of response patterns which since the end of the war have become known as the Guttman scale. It was Guttman's great merit to have formulated clearly the problem of analyzing qualitative data by intrinsic criteria. The basic equations he developed for this purpose do not assume any underlying model or latent structure. All of the solutions he found have, therefore, an interest of their own which would not be affected by the present approach.

However, the actual response patterns which Guttman has studied would also find their place in a systematic survey of all possible latent structures; the perfect Guttman scale itself is especially easily located. It is, in the terms of the present approach, a latent distance scale with $d = 0$. Its latent structure is given in Table 17. (See Section III, paragraph 5, and Section V, paragraph 3.)

TABLE 17

LATENT STRUCTURE OF THE PERFECT GUTTMAN SCALE

Latent class frequencies	*Latent marginals for items* 1	2	3	4	...
n_{I}	1.00	1.00	1.00	1.00	...
n_{II}	0	1.00	1.00	1.00	...
n_{III}	0	0	1.00	1.00	...
n_{IV}	0	0	0	1.00	...
n_{V}	0	0	0	0	...
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	

The response pattern which would be derived from such a latent structure is given for three items in Table 18. We see that four scale patterns clearly emerge. The cross-product matrix has rank m, corresponding to the number of items. It has the interesting property that the tetrads formed on one side only of the main diagonal will all vanish.

The trace lines of a Guttman scale are of a peculiar form. They

are straight lines identical with the x-axis up to a certain point and then in a discontinuous jump become straight lines with an ordinate of 1. Algebraically they would have to be represented by polynomials of infinitely high degree. The corresponding latent class structure for a Guttman scale, therefore, is one with infinitely many classes, if one had available what Guttman calls the "full universe of items." With a limited number of such items, the number of

TABLE 18

RESPONSE PATTERNS IN THE PERFECT GUTTMAN SCALE (3 ITEMS)

Response pattern	*Number in latent class* *I*	*II*	*III*	*IV*	*Totals*
+ + +	n_{I}	0	0	0	n_{I}
− + +	0	n_{II}	0	0	n_{II}
+ − +	0	0	0	0	0
+ + −	0	0	0	0	0
+ − −	0	0	0	0	0
− + −	0	0	0	0	0
− − +	0	0	n_{III}	0	n_{III}
− − −	0	0	0	n_{IV}	n_{IV}

classes which can be formed is one higher than the number of items. The latent class frequencies are immediately related to the manifest marginals of a Guttman scale by the following series of equations:

$$(6.1)\quad p_1 = \frac{n_{\text{I}}}{n}, \qquad p_2 = \frac{n_{\text{I}} + n_{\text{II}}}{n}, \qquad p_3 = \frac{n_{\text{I}} + n_{\text{II}} + n_{\text{III}}}{n}, \qquad \text{etc.}$$

The perfect Guttman scale has perfect precision in the sense of the previous section. This can immediately be seen from Table 18. The respondents who give one of the four admissible response patterns recruit themselves uniquely from one of the four latent classes. In a perfect scale, therefore, probability considerations do not enter.

As Guttman points out, empirical data will never furnish a perfect scale in his sense. To measure the degree of perfection, he has introduced the notion of reproducibility. It is important to realize that this corresponds to the notion of "fit" and not to the notion of "precision" in latent structure analysis. The structures we have discussed in the previous two sections, even if they are mathematically perfect, always misclassify a number of people as to their latent class position: they all have a precision of less than 100 per cent. In addition, they have imperfect fit; this means that the actual

ultimate frequencies are somewhat different from those which would be generated by the corresponding latent structure. The perfect Guttman scale would misclassify no one, but, of course, an empirical Guttman scale has a certain misfit. If it were perfectly fitted, no one would give a response pattern other than the main scale types.

It is possible to compare the fit of any latent structure with that of any other. We can always compute fitted ultimate frequencies and compare them with the actual ones. But this cannot be done with Guttman's coefficient of reproducibility, which can be applied only to the deviation of fallible data from the latent structure of a perfect Guttman scale. Take as an example the latent distance case which we have studied in Section V, paragraph 3. The d-values of the four items are, on the average, less than 10 per cent. The fit is very close as can be seen from the comparison of the last two columns in Table 13. The discrepancy between the actual and the "fitted" ultimate frequencies is 50.8 cases. Suppose we desired to find how well the same actual data would fit a perfect Guttman scale. Table 19 is a recasting of Table 13 into terms of the Guttman scalogram. Errors in Guttman's sense are boldfaced. To obtain the coefficient of reproducibility, the errors are summed and divided by the product of the sample size and the number of items in the test; this quotient is then subtracted from 1.00. In this case the coefficient of reproducibility is 94.0 per cent, a fairly high value. Now, if we apply our standard notion of misfit to a Guttman scale, we have to compute the "fitted" ultimate frequencies which a perfect Guttman scale with a given set of manifest marginals would generate. From equations (6.1) we get the latent subclasses. Inserting them in Table 18 (extended to four items), we get the full latent structure. With the usual procedure we get the fitted ultimate frequencies which were inserted as the last column in Table 19. The difference between the last two columns gives a measure of fit, which makes the Guttman scale comparable to other latent structures. The discrepancy is 532 cases out of 1,000.

In turn, if one were to make a χ^2 test of goodness of fit, one would have to remember that the perfect Guttman scale with 4 items has only 5 parameters—the 5 latent class frequencies—while the latent distance scale has 9 parameters, one more for each item. Therefore, a higher misfit in the Guttman scale can be accounted for by chance. It also should be remembered that the precision of a Guttman scale is always perfect.

Guttman has also introduced the notion of a quasi scale. While he has not yet given it a precise mathematical formulation, it is reasonable to assume that he would consider all the structures which have latent linearity as quasi scales. In this sense, one might say that the bulk of the work done in latent structure analysis is devoted to the classification and study of quasi scales.

The scalogram procedure by which Guttman studies the suitability of an item for a Guttman scale can be described in terms of trace

TABLE 19

SCALOGRAM ANALYSIS OF LATENT DISTANCE DATA OF TABLE 13

	Scalogram									
	+				−					
Response pattern	*1*	*2*	*3*	*4*	*1*	*2*	*3*	*4*	*Actual total*	*Fitted total*
+ + + +	X	X	X	X					75	138
+ + − +	X	X		X			X		10	0
+ + + −	X	X	X					X	8	0
+ − + +	X		X	X		X			14	0
− + + +		X	X	X	X				110	169
− + + −		X	X		X			X	11	0
− − + +			X	X	X	X			141	124
− + − +		X		X	X		X		49	0
+ − − +	X			X		X	X		11	0
− − − +				X	X	X	X		161	140
+ + − −	X	X					X	X	3	0
+ − + −	X		X			X		X	8	0
− − + −			X		X	X		X	64	0
− + − −		X			X		X	X	41	0
+ − − −	X					X	X	X	9	0
− − − −					X	X	X	X	285	429

lines (an observation made to the writer by S. A. Stouffer). The respondents are ordered first with the help of a preliminary score, say, the number of items they endorse. This corresponds (though very crudely, perhaps) to classifying respondents according to their latent class position, in as much as this preliminary score is likely to correlate rather highly with the underlying continuum. Then, for a Guttman item, the following criterion is applied: up to a certain value of the score almost everyone should endorse the item. Beyond this point, presumably, no one should endorse it. In other words, the item should show the type of discontinuous trace line previously described. Now nothing prevents one from using the scalogram technique to approximate trace lines of *other* types. One might request, for instance, that the proportion of people endorsing

an item should go down fairly proportionately with the progress of the general score. Or, for a middle item in a Thurstone scale, one would request that the endorsements be highest in the middle and go continuously down toward the two ends of the general score. Stouffer's recently developed idea has been applied to negatively and positively worded questions included in the same test. The scalogram lines show clearly the concave and convex forms required by general trace line considerations. Further work along this line has been started.

A final remark can be made regarding the type of items which are likely to fit into a Guttman scale. Let us first distinguish between two idealized extreme types of questions, which we shall call "projective" items and "yardstick" items. The essence of the first type is that the respondent is not aware of the purpose of the question. Suppose we show pictures of two men which look reasonably alike and give the one an Italian and the other an Anglo-Saxon name. Then, let us ask a sample of people which of the two looks nicer. If we find certain groups of respondents who consistently dislike the man with the Italian name, we will usually accept such a question as a useful test of prejudice. We would expect the trace line of such an item to be a rather smoothly ascending curve. The more prejudiced a person is, the more probable is his dislike for the picture of the man with the Italian name. Take, on the other hand, as an example of a "yardstick" item, a question which contains a term such as "Do you dislike very much . . .?" or any other term which connotes explicitly a position along the attitude continuum. The trace line of such a question is likely to make a sharp ascent close to the point of the underlying continuum which corresponds to the term used. Think, for instance, of a test designed to order respondents on a capitalistic-socialistic continuum. Suppose that the following two questions are included:

1. All economic systems have their advantages and disadvantages.
______ Agree ______ Disagree

2. Only a fool can have an opinion on a whole economic system.
______ Agree ______ Disagree

Both of these questions are likely to have bell-shaped trace lines; that is to say, they will be most frequently endorsed in the middle of the underlying continuum. But we would expect the second to have a sharper peak than the first (preliminary experiments have

shown that this assumption can actually be verified by the study of joint occurrence patterns). It is reasonable to expect that items fitting into a Guttman scale will be very much of the "yardstick" type. As a matter of fact, the mathematical conditions for a perfect Guttman scale are probably equivalent to the requirement that the admissible items do not contain any "projective" element.

In addition to this nonprojective character of the wording, Guttman items also have to be cumulative. Bell-shaped trace lines are obviously excluded. It might be mentioned that, as a consequence, the cross-product matrix of a Guttman scale can always be made to contain only positive values. This is not the case for the cross-product matrix of a Thurstone scale because middle items are admissible.

2. THE "TRACE LINE SCORES" OF RESPONSE PATTERNS

In previous pages, occasional remarks have been made about the scoring of response patterns. It remains now to discuss more systematically the relation between latent structure analysis and the scoring of tests the latent structure of which has been investigated. It is not the purpose of the present chapter to carry such a task out in full; this is left for a subsequent text in preparation for the RAND Corporation. But the preceding discussion would remain too incomplete if some of the main considerations were not summarized.

It will be remembered from Section I that in the general model there corresponds a trace line $y = f_i(x)$ to each item i. To the joint occurrence of items, say, i and j, there corresponds a trace line $y = f_{ij}(x)$. It is quite obvious from the previous discussion that every response pattern has its trace line. The response patterns $(+\ -)$ and $(-\ +)$, for instance, would have respectively the trace lines:

$$y = f_i(x)\,[1 - f_j(x)]$$

and

$$y = f_j(x)\,[1 - f_i(x)]$$

The meaning of these trace lines should be quite clear by now: they give for every point on the underlying x-continuum the probability that just this response pattern can be expected; or, in other words, the proportion of people who give this response pattern among all those who have the given x-value. It is easy to generalize the discussion of Section IV, paragraph 5, in terms of trace lines. Assume

for a moment that the trace lines are fully known. If a person then gives a certain response pattern, we still do not know his precise position on the underlying continuum. But we know for each point on the x-axis *how great the probability is that such a respondent has his position at that point.* And this is all we can ever know from the point of view of the present theory. Any score we might develop would only summarize this kind of probability distribution. To illustrate this idea we will address ourselves to the simplest but most basic scoring problem conceivable: how to decide which of the two response patterns $(+\ -)$ and $(-\ +)$ should have the higher score.

To be numerically concrete, we shall assume that a latent structure was found with linear trace lines. The latent class frequencies were: $n_{\mathrm{I}} = 600$, $n_{\mathrm{II}} = 400$. Two items in the test showed the following latent marginals:

TABLE 20

SCHEMATIC EXAMPLE FOR ORDERING RESPONSE PATTERNS

Latent class frequencies	*Latent marginals for items* 1	2
$n_{\mathrm{I}} = 600$	.90	.70
$n_{\mathrm{II}} = 400$	.20	.30

We know then from Section III, paragraph 2, that the manifest marginals of the two items are:

$$p_1 = \frac{n_1}{n} = .60 \times .90 + .40 \times .20 = .62$$

$$p_2 = \frac{n_2}{n} = .60 \times .70 + .40 \times .30 = .54$$

We also know from the same discussion that

$$S_1 = (.90 - .20)\sqrt{\frac{600 \times 400}{(1000)^2}} = .34$$

and

$$S_2 = (.70 - .30)\sqrt{\frac{600 \times 400}{(1000)^2}} = .20$$

Now we shall actually draw the linear trace lines for these two items. From Section I, paragraph 5, we know that in the linear case the slope of the trace lines is S, and their intersection with the

y-axis p_i, provided we choose a frame of reference so that its origin is at the mean of the distribution curve and the units on the x-axis are so chosen that $\sigma = 1$.

In this discussion we also need the trace lines for the negative replies to an item, which obviously have the equations:

$$y = 1 - (p_i + S_i x) = q_i - S_i x$$

where $q_i = 1 - p_i$.

In Figure 2a we draw the positive trace line for item (1) and the negative for item (2), corresponding to the pattern $(+\ -)$. Added is the trace line for $(+\ -)$ which is

$$y = (.62 + .34x)(.46 - .20x) = .29 + .04x - .07x^2$$

In Figure 2b we draw the negative trace line for item (1) and the positive for item (2), corresponding to the pattern $(-\ +)$, the trace line of which is also inserted. Its equation is

$$y = (.38 - .34x)(.54 + .20x) = .21 - .11x - .07x^2$$

Now we shall raise the question which is at the core of all scoring problems. Given two items like the ones in the present example, no one will doubt that $(+\ +)$ is the most positive reply and $(-\ -)$ the most negative. But in what order should $(+\ -)$ and $(-\ +)$ be inserted between the two? Is $(+\ -)$ more or less positive than $(-\ +)$? The intuitive inspection of Figures 2a and 2b will furnish the strong suggestion that $(+\ -)$ (2a) should have the higher, more positive, score. Its trace line lies more to the positive side than the trace line of $(-\ +)$ (2b).

There are a variety of ways in which this intuitive notion of the position of the trace lines could be formalized into an index. Two of them deserve special attention: the maximum probability of the response pattern and the expected value of a respondent's position.

To compare the highest probabilities of our example in a general way, we have to compare the position of the maxima in the following two functions:

$$(6.2a) \qquad (+\ -): \quad y = p_1 q_2 - x(p_1 S_2 - q_2 S_1) - x^2 S_1 S_2$$

$$(6.2b) \qquad (-\ +): \quad y = p_2 q_1 - x(p_2 S_1 - q_1 S_2) - x^2 S_1 S_2$$

The roots of the first derivatives are respectively:

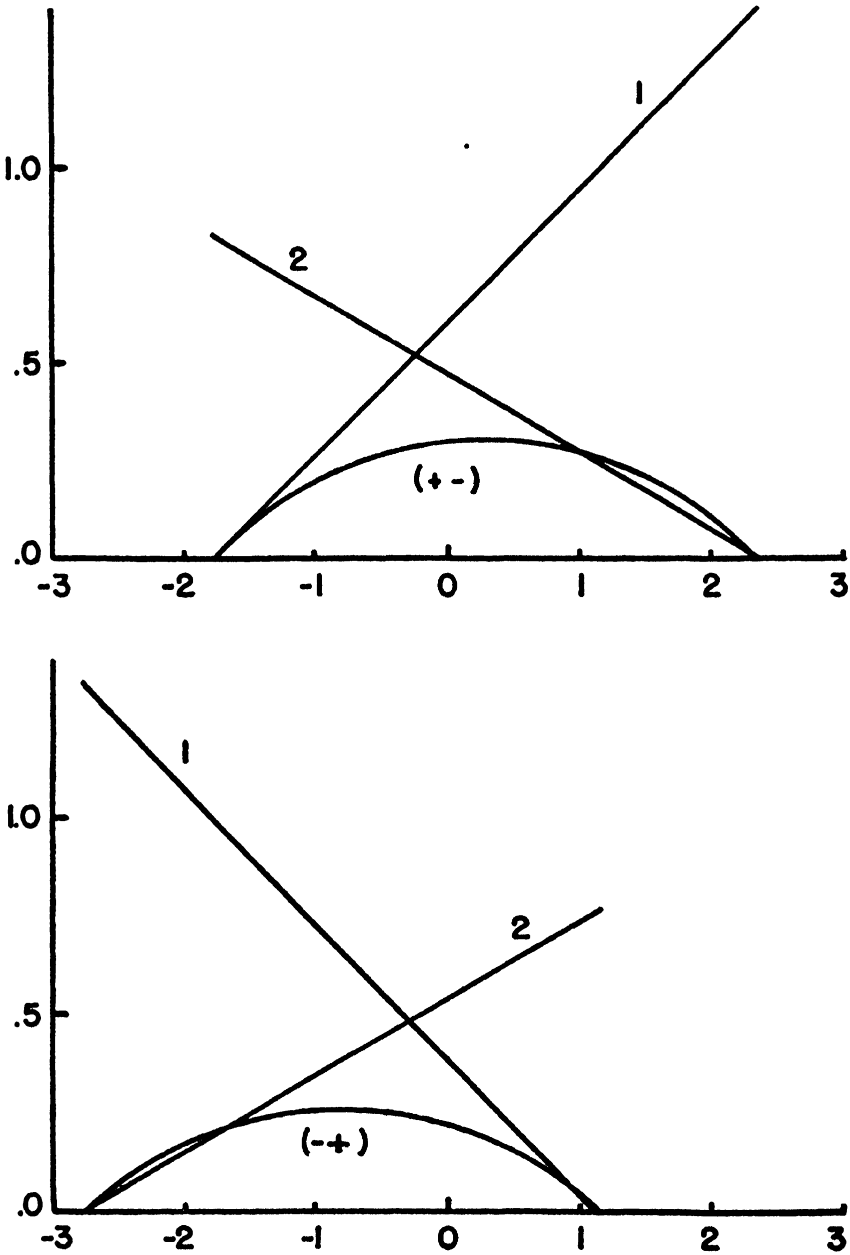


Figure 2a (top) and Figure 2b (bottom).

$$
\begin{aligned}
x \max \text{ for } (+\ -) &= \frac{-p_1}{2S_1} + \frac{q_2}{2S_2} \\
x \max \text{ for } (-\ +) &= \frac{-p_2}{2S_2} + \frac{q_1}{2S_1}
\end{aligned} \tag{6.3}
$$

This we shall call the *maximum probability score* (MPS). The relative position of $(+\ -)$ or $(-\ +)$ depends on the sign of the difference:

$$\frac{(q_2S_1 - p_1S_2)}{2S_1S_2} - \frac{(q_1S_2 - p_2S_1)}{2S_1S_2}$$

Remembering that $q = 1 - p$, this difference equals $\dfrac{(S_1 - S_2)}{2S_1S_2}$. We therefore have the result that $(+\ -)$ will have a higher MPS than $(-\ +)$ if $S_1 > S_2$. It can easily be seen that this result can be generalized: if $S_1 > S_2$, then any response pattern $(\cdot\ \cdot\ \cdot + -)$ will have a higher score than the corresponding pattern $(\cdot\ \cdot\ \cdot - +)$. More complex comparisons like, e.g., $(+ + - -)$ versus $(- - + +)$ can be decided by a similar reasoning.

Now, let us see what happens if the expected value is used as a score. People are distributed along the x-axis according to a distribution curve, $\phi(x)$. If, therefore, a respondent is found to give a response pattern with the trace line, say, $f_{12}(x)$, the probability that he comes from a position x is $f_{12}(x)\phi(x)$. This is a case where the classical theory of inverse probability applies. The respondent's expected position, therefore, is

$$\int_{-\infty}^{+\infty} xf_{12}(x)\phi(x)dx \tag{6.4}$$

Calling this the *expected value score* (EVS) and applying it to (6.2a), we obtain, e.g., for $(+\ -)$

$$p_1q_2M^{(1)} - (p_1S_2 - q_2S_1)\,M^{(2)} - S_1S_2M^{(3)}$$

In our frame of reference, $M^{(1)} = 0$ and $M^{(2)} = 1$. $M^{(3)}$, in this case, as we know from (1.8), is K, the computation of which was discussed in detail in Section IV (see especially paragraph 3). The two expected values are therefore:

$$\text{(EVS) for } (+\ -) = S_1S_2\left(\frac{q_2}{S_2} - \frac{p_1}{S_1} - K\right)$$

$$\text{(EVS) for } (-\ +) = S_1S_2\left(\frac{q_1}{S_1} - \frac{p_2}{S_2} - K\right)$$

The difference between these two score values is $(S_1 - S_2)$ which is essentially the same as for the (MPS) values previously derived (6.3). With linear trace lines, then, both types of scores would lead to the same ordering of response patterns. With more items and with higher degree trace lines, the relationships and computations become considerably more complicated. But the basic approach to scoring remains the same: no notion of "position of items" or "position of response patterns" is introduced into the model: it provides only for the notion of "position of people." *The score of a response pattern is the average position of the people who have answered the test in this specific way*, expressed as an (MPS) or an (EVS) or any other average.

It is clear that not only response patterns but also single items have scores. If the only information we have about a respondent is that he has answered item i positively we would have to put his expected position at $\int_{-\infty}^{+\infty} xf_i(x)\phi(x)dx$, which turns out to be $p_iM^{(1)} + S_i\sigma$ with linear trace lines or in the simplified frame of reference just S_i, a value which played such a prominent role in all computations of Sections I and II. With higher degree trace lines similar values can be derived.

It is worth noticing that one could in the same way derive more complicated indices, e.g., the dispersion of a response pattern measured by $\int_{-\infty}^{+\infty} x^2f(x)\phi(x)dx$, and so on.

3. LATENT CLASSIFICATIONS AND TEST CONSTRUCTION

We have thus a precise definition of a "score" in terms of a trace line model. In case we can derive the distribution function $\phi(x)$ from manifest data we then can carry out all necessary computation. This has been done successfully for the linear case. We know that K is the skewness of $\phi(x)$. Working with four items we can derive the following formula:[8]

[8] In the present paragraph, the symbol μ_k refers to the traditional statistical definition of the kth moment about the mean; it is not to be confused with the definition of a link, μ_i, used elsewhere in Chapters 10 and 11. The symbol $M^{(k)}$ refers to the kth moment about the origin.

$$\frac{\Delta_{ij.kl}}{[ij]\,[kl]\,p_{kl}} + K^2 + 1 = \frac{\mu_4}{\sigma^4} = \beta_2 \tag{6.5}$$

where

$$\Delta_{ij.kl} = [ij;kl] - \frac{[ij;k]\,[ij;l]}{[ij]}$$

This means that we then also know the kurtosis of $\phi(x)$. By adding terms we can derive higher and higher moments of $\phi(x)$ and therefore compute the expected value score for more complex response patterns; and only *origin and unit are arbitrary* in those scores.

If we set

$$[ij;kl] = \frac{[ij;k]\,[ij;l]}{[ij]} \tag{6.6}$$

the relationship (6.5) simplifies to $\beta_2 = K^2 + 1$. The identity (6.6) is not new to the reader of Chapter 10. It is the third condition necessary for the reducibility to two latent subclasses (see Section II, paragraph 7). The corresponding relationship between the third and fourth moments about the mean is not new to the mathematician. Take the following distribution function: the population is concentrated at two points: $x_{\text{II}} = -\sqrt{\frac{\nu_{\text{I}}}{\nu_{\text{II}}}}$ and $x_{\text{I}} = \sqrt{\frac{\nu_{\text{II}}}{\nu_{\text{I}}}}$; there are ν_{II} and ν_{I} people at the two points respectively, with $\nu_{\text{I}} + \nu_{\text{II}} = 1$. Such a distribution has a mean of zero, a standard deviation of 1 and $\mu_4 = \mu_3^2 + 1$.

This example permits us to see more clearly the relation between the general trace line model and the equivalent latent class structure.[9] The latter turns out to be a special case of the former, originating from a specific assumption about the distribution curve $\phi(x)$. Obviously other assumptions can be made, or none at all. The discussion of scoring in the linear case, carried out in paragraph 2 of this section, was valid for any kind of $\phi(x)$. The consequences of various assumptions about $\phi(x)$ and their relation to manifest data are now under investigation. The assumption of a discontinuous concentration of the population at a few points—mathematically equivalent to the existence of latent subclasses—has special bearing on test construction and needs some further discussion.

[9] The writer wishes to thank Professors Mosteller and Savage for helpful discussions on this point.

In attitude studies we are not only interested in minute differences between persons as we would be, e.g., in aptitude measurement. Volumes I and II of the present series use large numbers of attitude tests, which rarely if ever classify respondents into more than a few groups. This is probably due not merely to lack of precise instruments of measurement. A reflection upon the nature of attitudes shows that it is often quite realistic to think in terms of a few broad classes. People certainly vary according to whether they are radical or conservative, whether they are prejudiced toward Negroes or not; and from day to day, from one mood to the next and from one social situation to the next, their attitudes usually vary greatly.[10] These variations, however, are likely to stay within certain bounds: a soldier friendly to Negroes might occasionally be ruder to a Negro than to a white man, but he will not easily participate in a lynching. These observations are not meant to start a precise analysis; they merely should suggest that there is nothing disquieting in thinking of broader classes of attitudes, "natural" segments in an attitude continuum.

It is, of course, possible to achieve such broader classifications by combining the scores of any attitude test in arbitrary intervals. But the present approach seems to indicate that a given set of empirical data has, so to speak, its *most appropriate segmentalization.* It is the one which first assumes the degree of trace lines indicated by the rank conditions of the joint occurrences of second and third order and then "forces" the higher order and especially the ultimate frequencies into the "nearest" latent class structure.

The procedure for a concrete study as suggested by our analysis would then be as follows: the first question to raise is how many attitude groups are needed to classify our respondents. Suppose we decide on four groups. Then, we will want to pick such test items as are likely to have polynomial trace lines of third degree. It was mentioned before that there is a connection between the wording of questions and their probable trace lines. A considerable amount of experimental work will be necessary to establish this connection so clearly that it can be used in daily practice. This writer, however, can testify that it is not too difficult to make a pretty good guess as to the rank condition of response patterns from a careful study of question content. The final decision, of course, in every concrete test construction will have to be made in

[10] Evidence on this point will be published in a monograph by Patricia Kendall on the use of repeated interviews in the study of attitudes.

the light of actual data and therefore one will always want to start with more items than are finally needed in the test. Once a test has been found which satisfies the conditions for reducibility to four latent subclasses, the next step is implied in our analysis. We can compute a latent structure and put people into that one of the four classes into which they most probably belong, according to their response patterns.

This procedure is likely to be greatly facilitated by a theorem which Louis Guttman has developed.[11] According to this theorem, the frequencies of response patterns with bimodal latent structure tend to become smaller as additional items of the same rank structure are added. In response patterns with unimodal latent structure, the highest probability tends toward 1 and the other probabilities toward 0 as more items are added. This means that it is possible to get as near to a precise latent classification as is desirable for the problem at hand. Or, to put it in other words, with increased number of items, the expected position of most of the respondents will move ever closer to one of the concentration points of $\phi(x)$. The expected value score will indicate which point each response pattern is likely to approach. The simplest way to form the score in this case is to assign to the ordered set of latent classes a convenient series of weights and thus compute the averages of the latent probabilities.

The discussion on scoring developed in paragraph 2 above thus applies to the equivalent latent structure as well as to the general trace line model. As long as our tests have only a limited number of items, the question will remain as to which scoring system is preferable. *If a specific set of fallible data makes it equally reasonable to fit a latent class structure or a continuous distribution curve, which would be more advantageous?* In this writer's opinion, the answer would be about as follows:

If the material indicates the existence of linear trace lines, then much speaks in favor of a continuous model. The distribution curve can actually be computed; the scores of the response patterns are unique and permit for all practical purposes a metrical interpretation; the relationships to other variables can be studied more carefully; comparisons between groups and between interviews taken at different times can be made in considerable detail. It is also easier to find items which satisfy the two conditions for linear

[11] From a written communication by Professor Guttman.

trace lines if one does not also require the additional condition for the existence of a latent dichotomy.

But there are many research situations where linear trace lines are not desirable nor feasible. The former will be the case if "neutral items" are used of the type characteristic of a Thurstone scale. Linear trace lines are unlikely to exist when large groups of "don't know" answers occur in the manifest material; or when the "natural" cut of the dichotomous questions shows a great variation. Once the material requires nonlinear trace lines, then it is undoubtedly better to assume, wherever the data permit, a discontinuous $\phi(x)$. The computations are easier and the possible interpretations of latent classes are more helpful than they are with distribution curves which have some of their moments (in addition to the mean and σ) undetermined.

Once the assumption of a discontinuous $\phi(x)$ has been made, it is, strictly speaking, no longer legitimate to use an expected value score (EVS) for each separate response pattern as we did in connection with our example in paragraph 2 above. For then one puts the expected position of a respondent on a point where he could not be, according to the discontinuous assumption. But it seems that an EVS in this case would lead to an ordering rather similar to the one which would have been derived from the continuous assumption and with much less computational labor. Therefore, our EVS in the case of a latent class structure analysis seems permissible as a substitute device. But the most adequate use of the latent structure is to put the respondents into the latent class where they most probably belong, perhaps leaving unclassified the cases with very similar latent probabilities in more than one latent subclass.

4. RELATION OF LATENT STRUCTURE ANALYSIS TO FACTOR ANALYSIS

The present theory has a number of similarities to and differences from factor analysis.

The greatest similarity is in regard to logic. Much of what Thurstone in his latest book has said about the role of mathematical models applies to both approaches. As a matter of fact, the present system has developed from a desire to adapt factor analysis to qualitative material.

Previous writers have applied factor analysis to dichotomies in the following way: fourfold tables were formed between all the

pairs of items and then so-called correlation coefficients were attached to these fourfold tables. Once they were accepted and entered into a correlation matrix, the whole computing mechanism of factor analysis could of course be applied. But this seems to be an unjustifiable procedure, as can easily be seen by the fact that different results were obtained according to whether one worked with tetrachoric coefficients or the Pearson point correlations.[12] The present approach avoids this difficulty by using the formulations of factor analysis in such a form that the notion of correlation is never needed. The only concept which is needed mathematically is the notion of independence of variables, which has a unique meaning for dichotomies.

As a mathematical result the whole theory can be based on the analysis of joint occurrence matrices. But as has been seen repeatedly, it is possible to base it also on a good analogy to the correlation matrix: the cross-product matrix. This would make it appear that the present theory is nothing else than factor analysis where the form $[ij]$ plays the role of a correlation coefficient.

But this is not so, as can be shown in the simplest case of one factor. If in a system of tests all pairs of tests have a partial correlation of zero when one factor is partialled out, then the correlation matrix has a rank of one. How about a system of dichotomies, the interrelationships of which can be explained by one latent characteristic? The $[ij]$ matrix has a rank of 1 if the latent characteristic is a dichotomy. It has a rank of 2 if the latent characteristic is a trichotomy, etc. Why does not segmentalization of the factor appear in factor analysis?

As far as this writer can see at the moment, this is due to the fact that factor analysis assumes a linear relationship between the factor and the test. This would correspond to a fully degenerate latent structure with rank 1 for the ensuing $[ij]$ matrix. If curvilinear relationships between the test and the factor are admitted, then even one factor might lead to correlation matrices with higher rank. The analogy here is with polynomial trace lines of higher degree. The latent structure of a latent trichotomy, e.g., may show a

[12] See George A. Ferguson, "The Factorial Interpretation of Test Difficulty," *Psychometrika*, Vol. 6, No. 5 (October 1941), pp. 323–329; and R. J. Wherry and R. H. Gaylord, "Factor Pattern of Test Items and Tests as a Function of the Correlation Coefficient: Content, Difficulty, and Content Error Factors," *Psychometrika*, Vol. 9, No. 4 (December 1944), pp. 237–244, for discussion of the difficulties ensuing from such a procedure.

curvilinear relationship between the item and the underlying continuum.

The matter can be put in still another form. Factor analysis deals only with zero-order correlations. In Pearson's correlation theory, the higher-order correlations do not add any new information; they can be computed from the zero-order correlations. In the present approach, M-matrices of higher signature are used. Joint occurrences of three or more items are empirical data and not mathematical derivatives from joint occurrences between any two items.

The following question may be raised and offered for further investigation: suppose a factor analyst has obtained the score values of a number of tests. Instead of computing the correlation coefficients between the scores, he could dichotomize all his test results so that he would end up with just fourfold tables between the tests. He would feel that he had thrown away a great deal of information. But has he really? Suppose now he does not proceed to tack a correlation coefficient upon these fourfold tables but proceeds to tabulate the higher-order joint occurrences of his dichotomies up to the point where he can carry on a latent structure analysis. Where would he come out? He would know in the end the latent marginals of his tests which would approximately correspond to factor loadings. And, if his latent structure has three or more subclasses, he would have a pretty good idea as to the actual shape of the regression curve between test and factor. It should be remembered, however, that for such larger λ, equations of very high order would have to be solved, which would entail a great deal of work. As a matter of fact, as was pointed out before, a way to solve these equations generally has not yet been developed. But in the age of electronic computing machines, it should be possible for students competent in this field to tackle the basic matrix equations and to furnish us with manageable procedures.

It seems then that for dichotomies the present approach forms the appropriate parallel to factor analysis. To suggest that it also be applied to quantitative material is at present a mere speculation. But if some examples were carried out on material which has been factorized before, clarification might be gained. But one need not be restricted to empirical studies. The following mathematical extension of the present approach deserves serious attention.

Suppose the manifest data are not given as dichotomies but as

trichotomies. In Section III, paragraph 3, we discussed what happens if the 3 steps are "collapsed" to a dichotomy. We could also keep them separate. For any two items, this would give a "ninefold" instead of a fourfold table. Each of these 9 combinations would have a trace line. We could set up accounting equations just as we did in Section I, only for m trichotomous items we would now have 3^m such relationships between manifest and latent data. It can be easily shown that all the algebra developed in the preceding chapter can be directly applied. Suppose we speak in terms of answers "Yes," "No," and "Don't know." First, we would consider the answer "Yes" as a positive occurrence and compute the corresponding latent marginals. Then the same would be done for the answer "No." The latent marginals for the answer "Don't know" would come out by subtracting the first two sets from a total of one. The amount of labor would be greatly increased, especially because additional adjustments would become necessary. Furthermore, it is very dubious whether much would be gained by this complication; but no new theoretical element or difficulty would be added.

Now, extend this idea not to 3 steps for each item, but to many, say, w_i for item i. Then we would be in the situation of "continuous" variables with w_i intervals. Each combination of test scores would furnish one accounting equation. At this point a reformulation of the basic algebra would be necessary, and no work has yet been done in this direction. But this extension, if it turns out to be mathematically feasible, would lead to a general factor analysis of which the Thurstone tradition would form the continuous linear case; latent structure analysis would be its specific application to dichotomous data without restriction as to linearity. The latter approach, restricted as it is, still has many tasks waiting for the mathematician as well as for the empirically minded social scientist.

CHAPTER 12

TWO CASE STUDIES IN PREDICTION: INTRODUCTORY COMMENTS[1]

ONE of the objectives of social science is to develop techniques for prediction which will enable the user to predict more efficiently than if these techniques did not exist. By more efficiently we mean that the percentage of correct predictions is higher, on the average, than would otherwise be obtained for the same expenditure of effort.

The Research Branch frequently in its surveys was in the position of having to make informal or implicit predictions. For example, the interpretation of any observed relationship between two variables usually carries with it the implication that the association found at a given time and place will be found at some future time and place. In some cases, especially when controlled experiments have been used as in the data constituting Volume III of this series, the results carry the additional implication that if one manipulates X, one can expect a change in Y—in other words, that X can be regarded now and in the future, as a "cause of Y." All too often to satisfy the scientific consciences of members of the Research Branch, predictions forced by military exigencies had to be made when the data were quite inadequate.

Various efforts at prediction have been mentioned in Volumes I and II. One of the more ambitious, and also one of the less scientific, was the elaborate forecast reported in Volume II, Chapter 12, of general psychological conditions to be expected with the defeat of Germany. Much of the content of the predictive memorandum quoted there in full deserves to be classified as "enlightened common sense" rather than as a forecast directly arising either from scientific principles or from data permitting actuarial induction. Yet some principles were used and available data were taken into account as far as possible—and, with exceptions duly noted in the chapter, the predictions hit the mark fairly well.

[1] By Samuel A. Stouffer.

Because the improvement of methods of prediction is so important in social science, it seemed worth while to present at some length in the present methodological volume two detailed case studies of prediction in which an explicit effort was made not just to forecast what certain general tendencies would emerge, but actually to specify the proportions of men who could be found later to behave in such and such a fashion.

First, in Chapters 13 and 14, by Shirley A. Star, we shall take the reader through the stages in the construction of a pencil and paper test for screening psychoneurotics. The final form of this test was used routinely in the last year of the war at all induction stations in the United States.

Second, in Chapters 15 and 16, by John A. Clausen, we shall examine the problems involved in predicting what kind of employment men would seek upon discharge from the Army, as well as how great the demand would be for further schooling.

These studies should be instructive for the student of social science. They are by no means models of clean and simple prediction. The complexities of the problems and the ambiguities of the criterion situations called for much technical ingenuity and, in the final analysis, for the exercise of some enlightened guesswork. In both instances, the Branch felt that the value of the results more than justified the effort. Many pitfalls were anticipated and successfully evaded, others were not. By a frank recording of these accounts of prediction, it is hoped that future students can help equip themselves to do analogous tasks better.

As is so often the case with problems of prediction, the major difficulty lay in the lack of clarity as to the criterion situation. If one is using a poll to predict a presidential election, one has a simple and definite criterion in the number of votes actually cast for each candidate at the presidential election. If one is predicting, however, on the basis of test scores, which men psychiatrists will reject for the Army after a psychiatric examination at induction stations, it turns out that the criteria are variable indeed.

In Chapters 13 and 14 a study is reported of over 100,000 men constituting the entire population examined at all the induction stations in the United States in August 1945. For each man we have data on his score on the Neuropsychiatric Screening Adjunct (NSA), i.e., the screening test developed by the Research Branch for the Surgeon General, and on the disposition made by the examining physicians. There we see a rather astonishing fact—that, although the test scores had much the same frequency distribution through-

out the United States, the proportions of men rejected on psychiatric grounds varied all the way from 0.5 per cent at one induction station to 50.6 per cent at another.

Not only did the psychiatric rejection rate fluctuate from one induction station to another, but so also did the reasons for psychiatric rejections. If, for each station with fifty or more psychiatric rejects, we set *all* psychiatric rejects equal to 100 per cent, we find that the proportions among these psychiatric rejects classified as *psychoneurotic* varied from 2.7 per cent to 90.2 per cent, while the proportions classified as *psychopathic* varied from nothing up to 81.3 per cent. Such variations may seem almost fantastic, but, of course, one must remember that psychiatry is still far from an exact science, that the number of competent psychiatrists is not too large, and that the time which physicians could devote to the examination of any one man was extremely limited.

Variations were especially large in the South, as might be expected due to lack of trained psychiatrists. Thus at the induction station in New Orleans, 74.6 per cent of the psychiatric rejects were found to be psychoneurotic, but only 22.7 per cent were at Oklahoma City. Only 2.2 per cent of the psychiatric rejects in New Orleans were found to be psychopathic, as compared with 70.9 per cent in Oklahoma City. But such variations were not confined to the South. In Boston, 57.9 per cent of the psychiatric rejects were found to be psychoneurotics; in New Haven only 25.0 per cent. In Chicago the figure was 71.8 per cent; in Detroit, 31.5 per cent. In San Francisco the figure was 69.1 per cent; in Seattle-Portland the figure was 22.9 per cent.

The scoring of the screening test was so designed that about a third of American young men, on the average, would manifest signs indicative of the need of very careful psychiatric examination. This does not mean, as is fully explained in the next two chapters, that all these men should have been rejected. Rather, the cutting point was set high enough to include among the one third practically all of the psychiatric cases, as well as some who, upon careful examination, should be accepted by the Army. Part of the prediction problem was to find whether or not this one third did include most of the eventual psychiatric rejects.

Actually, as is shown in detail in Chapter 14, this one third screened by the test included 80.8 per cent of the psychoneurotic rejects, 68.2 per cent of the psychopathic rejects, 70.8 per cent of the psychotic rejects, and 56.4 per cent of all other psychiatric rejects. All told, it screened 69.5 per cent of all psychiatric rejects.

At the same time, it screened, as cases also worth examining, 21.8 per cent of the men eventually admitted and 30.3 per cent of the men rejected by the physicians on other than psychiatric grounds.

In spite of the very great variability from one induction station to another in psychiatric diagnosis, there was no station with as many as fifty psychiatric rejects in which the test did not screen at least half of those eventually rejected for psychiatric grounds, and in all but four of these stations the proportions screened were between 60 per cent and 96 per cent.

The tables shown in Chapter 14 have, therefore, an importance transcending a mere exercise in methodology. The variability of psychiatric diagnosis is not exactly news to psychiatrists, but until this study was made there was no uniform scoring device against which the variability throughout the United States could be compared. Hence, if there was three times as high a psychiatric rejection rate in one city as in another, there was no answer if somebody chose to argue that maybe one city did have three times as many psychiatric cases in proportion to its population as another.

How the Research Branch data throw light on such discrepancies may be illustrated in detail by a comparison of data from the Chicago and Detroit induction stations:

	Chicago	*Detroit*	*Number of cases*
Proportion rejected on psychiatric grounds among all men examined	7.6%	21.6%	*4,523*
Proportion screened by NSA as worth careful examination	24.1	26.9	*4,235*

We see that, *in spite of the fact* that the test scores showed about the same proportion in the two populations as needing careful examination (24.1 per cent in Chicago and 26.9 per cent in Detroit), the Chicago psychiatric examination rejected only 7.6 per cent while the Detroit examination rejected nearly three times as many. Let us continue:

	Chicago	*Detroit*
Psychiatric diagnosis of all psychiatric rejects:		
Psychoneurotic	71.8%	31.5%
Psychopathic	10.7	39.0
Psychotic	1.2	—
All others	16.3	29.5
	100.0%	100.0%
Total number of rejects	*344*	*915*

Not only did Detroit reject three times as many as Chicago, but also among the rejects the "causes" of rejection were very different.

Because the distribution of test scores at the two stations was about the same, yet the psychiatric diagnoses were so different, one cannot expect to find the test very efficient in predicting psychiatrists' behavior. We have the following findings:

	Chicago	*Detroit*
Proportion screened by NSA as worth careful examination, among:		
Men later accepted for military service	19.5%	15.3%
All psychiatric rejects	64.6	60.9
Psychoneurotic rejects	69.0	88.8
Psychopathic rejects	62.2	58.2
Psychotic rejects	75.0	—
All other rejects	42.9	35.3

There can be no doubt that the test tended to screen a much larger proportion at each induction station among those subsequently rejected on psychiatric grounds than among those accepted. That the discriminating power of the test was as effective as these data show is all the more noteworthy in view of the widely divergent standards of psychiatric diagnosis, acceptance, and rejection which it seems probable were applied at the two stations.

The same kind of discrepancies, as between Chicago and Detroit, are also seen all over the United States, as the tables in Chapter 14 vividly portray. In general, the relationship of the NSA test scores to the psychiatric diagnosis was about the same elsewhere as at these two stations; on the average, somewhat more discriminating but not much more.

Thus we have illustrated a central problem of prediction in social science—namely, the difficulties of making predictions when the behavior to be predicted (in this case, actually, the behavior of psychiatrists with respect to the individual tested) is itself unstandardized and subject to great variation.

The second case study in prediction to be examined in detail in this volume also is one whose major difficulty lay in the differences between the test situation and the criterion situation. Here our problem was to predict what men would do after the war. It was of much practical importance to get this information as accurately as possible. Congress, in budgeting for the GI bill, needed a basis for some kind of estimate of the total cost of GI benefits for education. The Veterans Administration, Department of Agriculture,

and many other government agencies needed data for future planning. Various branches of the War Department were involved in vocational guidance and in service educational programs, and were particularly anxious to find out two facts: the extent to which interest in various types of vocational careers was manifested; and the extent to which this interest represented realism or mere wishful thinking.

To ask a man fighting in the Pacific two years before the end of the war what he expected to do when the war was over—a war which might last many years as far as anyone knew—involved considerable temerity. The asking itself was simple, but the framing of questions and interpretation of the answers was another matter.

In Volume II, Chapter 13, some of the findings of these studies have been reported. But the steps involved in arriving at sets of questions which it was hoped would be satisfactory represent a research saga worth recording for its methodological interest. Many of the men had only the vaguest of plans and some had alternative plans, of which one or more may have been quite unrealistic. Moreover, nobody knew what the labor market would be like after the war. Some economists were predicting a depression, with millions unemployed. Whether or not soldiers were to be discharged into a country with factories running full blast or into a country with millions of people idle obviously was a contingency which could not be ignored in making predictions.

Chapters 15 and 16 trace the stages of the development of questionnaires for the analysis of postwar plans and summarize the results of two special studies, made at the close of the war, in which soldiers who had filled out Research Branch questionnaires were followed up several months after discharge. The follow-up studies provided a check on the kinds of errors in prediction made for each individual, where errors were made.

In both the studies of psychiatric diagnosis and the studies of postwar plans will be found some interesting examples of the application of scalogram analysis described in earlier chapters of this volume. The psychiatric example is especially interesting in this respect, since, incidentally, it provides some unique data for social psychologists who are interested in theories of abnormal psychology. As a result of many conferences with psychiatrists, fifteen scales or quasi scales by the scalogram method were constructed to cover a wide variety of problems thought to be involved in psychiatric breakdowns in the Army. The areas represented by the fifteen

scales or quasi scales in rank order of their discriminating value, as between psychoneurotics and normals, are:

1. Psychosomatic complaints
2. Personal adjustment
3. Childhood symptoms
4. Soldier role
5. Childhood fears
6. Sociability
7. Oversensitivity
8. Participation in sports
9. Worrying
10. Childhood fighting behavior
11. Relations with parents
12. Identification with war
13. Childhood school adjustment
14. Mobility
15. Emancipation from parents

One of the most remarkable results was the fact that a single one of these fifteen areas, namely, a quasi scale on psychosomatic complaints, was almost as discriminating as between psychoneurotics and others as the entire battery of items covering fifteen areas. Because of this fact the final test forms developed could be reduced to only a few items.

It may be noted in passing that attempts made by psychologists in the Research Branch to use modified projective techniques like the group Rorschach proved comparatively unrewarding.[2] This is neither a reflection on the theory of projective testing or on the practical utility of individual projective tests. It should be observed that the military situation was one in which there was no disadvantage to a malingerer in giving the "neurotic" answer even if he suspected it to be such. In other situations, involving application for private employment, for example, the opposite might be true; and in such situations projective tests, in improved form one can hope, may be found to be indispensable.

The postwar plans research is particularly valuable in serving as a warning against blind and indiscriminate use of scales. In some respects scales, of the kind described in the earlier chapters of this volume on scalogram analysis, were helpful in ordering respondents along single continua, such that each rank group had a higher probability of carrying out the predicted behavior than the rank group immediately lower. However, there may be a temptation for future students to become so devoted to the application of scaling that they may fail in the vitally important task of analyzing fully

[2] Actually, the Harrower-Erickson multiple choice Rorschach test proved to have no greater discriminating power than a single check-list item like, "Are you ever troubled with nervousness?" The Research Branch experimented with a brief type of open-end response to the Rorschach slides, which was much more promising, but would have required more time than was available to bring it to a stage of practical utility.

the situation in which they want to apply their scales. A careful study of Chapters 15 and 16 should, therefore, serve as an instructive corrective—at least in the present early stages of the development of scale theory.

The research reported here required decisions both by the respondent and by the analysts on the basis of weighing a good many values and contingencies. Items not fitting into a scale needed to be taken into account in assessing the probable materialization of contingencies and the implications of responses to questions getting at such contingencies. For example, a respondent may have definitely and strongly wanted to go back to school, but felt that he might not go back if he were lucky enough to get a certain specified kind of job. Taking this response at its face value, the prediction problem involves estimating the chances of his actually getting such a job. It is quite clear that the latter estimation depends in turn upon a forecast of the employment situation in general, of the probable openings for such jobs in particular, and of the chance that the respondent would qualify for such a job if it were available. Only if we are prepared to make such guesses (which may involve procedures for which scaling has not as yet been adapted) can we assign such a man with confidence to a scale position with respect to his likelihood of returning to school. We may, it is true, fit him into a broad scale category comprising all those who would go if they cannot get the job they want. But we do this at a sacrifice to predictability, since some men in such a category may have a high probability of getting that alternative job and some may have a low probability.

Chapters 15 and 16 provide some concrete data on the advantages and limitations of various *types* of questions for eliciting needed information as to a man's future intentions.

For example, at one time it seemed wise to begin with a very general free-answer question like, "What do you expect to do after the war?" Such a question had the advantage of encouraging a free answer unstructured and unbiased by its format—the nondirective approach. But it was not particularly useful, because many men answered it in terms of marriage or in terms of where they expected to live or in terms of recreation rather than in terms of occupation. Somewhat more useful was a question of the following type: "What kind of work do you think you will do right after the war? (Write down the name of the job and describe it as fully as you can.)" This brought responses in terms of job plans,

but was biased against the reporting of plans for further schooling. When schooling also was included in the format of the question, however, it was found to bias the answers too much in the direction of further schooling. Moreover, the problem of coding the free answers on thousands of questionnaires from all over the world made the cost of processing such free answers exceedingly heavy. The free-answer question, however, was frequently helpful in resolving ambiguities in answers to more direct questions.

The free-answer question was usually followed by a set of explicit check-list questions such as:

> Do you think you actually will go back to work for the same employer (company, person, etc.) you worked for before you came into the Army?
> ______ Yes, I'm quite sure I will
> ______ I may, but I'm not sure
> ______ No, I don't think I will
> ______ I worked for myself before I came into the Army
> ______ I was not working before I came into the Army

Sometimes, several such questions in a given area were asked in order to get a scale score. These questions were very useful, but they contained a large element of suggestion and there were a good many respondents who checked "Yes, I'm quite sure I will" to apparently mutually incompatible questions. All answers, of course, could have been true if each represented a different period in future time, but, in many instances, as further querying showed, they often represented simultaneous *alternatives* or wishful thinking.

To get at such alternatives, explicit contingency questions were introduced, e.g., "If you could get a good job, but not the kind of job you want, would you go back to full-time school?" It turned out that such contingency questions were not always answered in literal terms of the contingencies offered and they presented considerable difficulties in analysis, although they too were often necessary to "peel off" unrealistic responses to the more direct questions.

Perhaps the simplest device for classifying individuals who were considering alternative courses of action was a check list of alternatives such as the following:

> Check the one thing you probably will do *first of all*—after you have gotten your discharge and taken a vacation.
> ______ I will probably work for the employer I was working for before I came into the Army
> ______ I will probably work for salary, wages, or commission for some other employer

_____ I will probably go to full-time school or college
_____ I will probably farm for myself or with my family
_____ I will probably go into business for myself
_____ I am undecided what I will do

Such a question was easy to tabulate, and it could be used to spot respondents who were basically undecided without having to do an elaborate coding job. But it was not free from bias, for there was a tendency to check prestige items, like going into business, rather than an item closer to actual expectations. Hence, the analysis of answers to individual items, like "Do you think you will do farming when you get out of the Army?" followed by a series of further detailed questions to establish, by scale analysis or otherwise, the realism of an affirmative answer, was usually a more efficient procedure for establishing predictive cutting points than the simple alternative check list.

The data in Chapters 15 and 16 permit one to study these various types of approaches in terms of the degree to which they actually predicted or failed to predict subsequent behavior. It so happened that the employment situation in private industry was excellent after the war, although various factors, including shortages of materials, inhibited the opening up of new small businesses. In spite of the good labor market, the incentives to continue education were high, as Congress successively increased the allowances in the GI bill. On the whole the predictions were quite successful—especially with respect to working for an employer in private industry, with respect to farming, and with respect to going back to school; somewhat less so with respect to setting up one's own business or getting government jobs. Some of the reasons for this differential effectiveness are brought out clearly in the analysis.

In the chapters which follow, primary emphasis has been placed on describing the methods used, and the content findings are treated as incidental. To describe the operations adequately, a full-sized book would be needed for each of these two case studies in prediction. The condensation required by limitations of the present volume has been achieved at some cost to the student, who will have to follow the description rather patiently on some of the pages which are closely written.

It should be emphasized that these chapters represent concrete illustrations of problems of prediction. Many more such bodies of illustrative materials need to be assembled before we shall be ready to enunciate rules or principles about the conditions under which

prediction is or is not most likely to be successful. Such an effort should not be crudely empirical but should flow from considerations arising out of maturing theories of learning.

John Dollard, while a consultant in the Research Branch, made an incisive attempt to systematize some preliminary thinking about this problem.[3] Dollard visualized the problem as a relationship between three classes of situations, which he called the origin, test, and criterion situations. The *origin* situation is one in which the thought or opinion sentence is first hit upon and rehearsed. In the *test* situation the social surveyor intervenes with questions which tap a stream of thought sentences or opinion sentences presumably orienting behavior of the person surveyed. The *criterion* situation is the dilemma around which prediction is centered and in which the respondent must take whatever rewards and punishments ensue upon his action.

Prediction is possible, as Dollard sees it, to the extent that there is some correspondence between these three situations. There is in all of our social conditioning from early childhood a strain to match verbal expression with overt behavior and also thoughts with overt behavior. Children are trained in rehearsing directions received from parents and policed to see that they follow these directions correctly. Even when a child is asked to solve an arithmetic problem or to spell a word, "he goes through a process of inner rehearsal which is a kind of prediction of what the explicit answer will be. If his predictions prove frequently to be invalid, the psychologist may lay on him the curse of Binet and state that he has a low IQ."

Dollard advanced, very modestly, some preliminary hypotheses as to the conditions which seem to favor matching the opinion in the test situation to action in the criterion situation. Prediction, he suggested, is favored *if:*

1. *In the test situation no relevant response element is disconnected from thought and speech.* If some relevant elements are disconnected (i.e., unconscious) these elements will remain at play in the criterion situation though not registered in the test situation. Example: a man is not aware of his jealous tendencies toward his wife because they have never been excited. He might think, upon

[3] See John Dollard, "Under What Conditions Do Opinions Predict Behavior?" *Public Opinion Quarterly*, Vol. 12, No. 4 (Winter 1948–49), pp. 623–632. This paper, originally prepared while the author was serving in the Research Branch, was read before a joint meeting of the Washington Statistical Society and the Washington chapter of the Institute of Mathematical Statistics, March 9, 1944.

questioning, that he would not be jealous should a rival appear, but might find himself painfully so in the event.

2. *If the sentences of the respondent cover all response elements relevant to the criterion situation.* Persons with poor verbal skills may find it difficult to forecast their behavior, even when no automatic repression exists as in the example above.

3. *If the connection between sentence responses and overt action responses has been strongly rewarded in previous learning.* In some persons, who have not been rewarded for thinking first and then acting, a split occurs between thought and action. They are dreamers who do not make their dreams come true, the "phantasy" types.

4. *If the test situation provides stimuli only to the verbal response relevant to the criterion dilemma.* The test situation should not be corrupted by extraneous threats or rewards. For example, a factory worker may deny that he intends to join a union out of fear that the interrogator represents the boss.

5. *If the origin situation corresponds closely to the criterion situation.* A man can best predict what he will do in a future situation if he has been in about the same situation before and thus knows what it's all about. In the polling-election situation prediction is highly favored. The question, "Whom will you vote for?" is about the same in the two situations. But, as in so much of the illustrative material in Chapters 15 and 16, the test situations and criterion situations may have no such simple correspondence.

6. *If no new origin dilemma intervenes between the test and criterion situation.* A new experience arising between the two situations can upset the prediction. For example, a Southerner coming North for the first time might have the opinion that he would never sit by a Negro on a bus; but once in the North a new learning dilemma occurs, since the Southerner does not want to appear strange or bigoted and his Southern penchants against breaking caste etiquette have been withdrawn for the time being. A new origin experience has come into being.

7. *If the test question explicitly presents the conflict, i.e., anticipations of rewards and punishments, of the criterion dilemma.* A man will better predict what he will do in a future dilemma if he is told exactly what this dilemma will be. This condition cautions particular care in evoking the criterion situation vividly and specifying exactly the behavior to be predicted.

More efforts of this kind, to systematize thinking about the conditions of successful predictability, are needed. It is hoped that

the experience recorded in chapters following will serve, along with other such documents, as source material for the development of theories of prediction which, in Dollard's language, will help "relate the data of the social survey to the framework of a more general science of human behavior."

CHAPTER 13

THE SCREENING OF PSYCHONEUROTICS IN THE ARMY: TECHNICAL DEVELOPMENT OF TESTS[1]

THIS chapter is in three parts. Part 1 traces the process by which the Neuropsychiatric Screening Adjunct, used at all induction stations in the last year of the war, was developed. Parts 2 and 3 are, in a sense, appendixes. In Part 2 each item used in the initial study in the Army to discriminate between patients in psychoneurotic wards in station hospitals and a cross section of the Army is presented, with the percentages of psychoneurotics and "normals" endorsing each response category. In Part 3 the technical procedure for weighting items is described in detail.

PART 1

THE LOGIC AND PROCEDURE LYING BEHIND THE NEUROPSYCHIATRIC SCREENING ADJUNCT

The General Concept of Screening

The theory of screening tests is not particularly new;[2] they are merely devices for cutting off a portion of a population in such a way that it contains a disproportionate number of some subpopulation to be detected. To put it in the concrete terms of psychiatric screening, we may, of course, assuming infallible psychiatric diagnosis, discover every psychoneurotic in a given population by indi-

[1] By Shirley A. Star. The initial studies of psychoneurotics were begun by Robin M. Williams, Jr., and Louis Guttman in cooperation with Major John Appel, representative of the Neuropsychiatric Consultants' Division of the Surgeon General's Office, who had requested that the research be done. After the first pretest, Shirley A. Star succeeded Dr. Williams on the team. Dr. Guttman contributed the mathematical techniques employed and discussed here. The empirical work and later validation studies were under the direction of Miss Star.

[2] See W. A. Hunt, C. L. Wittson, and H. I. Harris, "The Screen Test in Military Selection," *Psychological Review*, Vol. 51, No. 1 (January 1944), pp. 37–46.

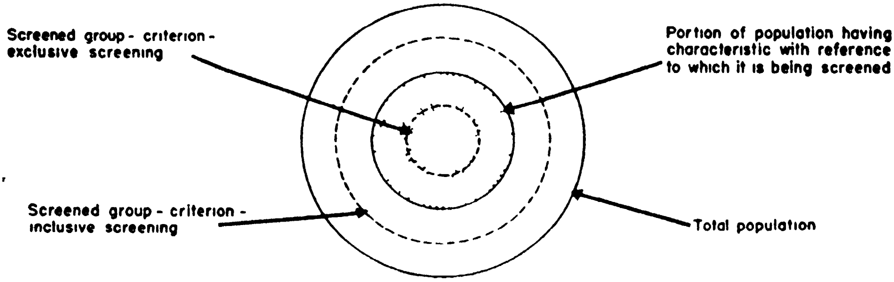
vidual psychiatric evaluation of each member of that population. On the other hand, this is a long and costly procedure, and often the number of psychiatrists needed for it is not available. A psychoneurotic screening test aims to economize psychiatrists' time by reducing the number who must be individually examined without a proportional reduction in the number of psychoneurotics who are examined.

From this standpoint, any stimulus to which there is a significant difference in response as between psychoneurotics and nonpsychoneurotics may be used as the basis of a screening test.[3] If the responses are suitably scored, and the men to receive psychiatric examinations are selected by reference to their scores, then the group examined by psychiatrists will contain a higher proportion of psychoneurotics than the population as a whole. Where the psychiatric facilities permit an individual examination of only a portion of the population, then selection of that portion by means of a screening test brings greater returns than random selection. Just what proportion of the population is to be considered screened by the test and referred for psychiatric evaluation is, where a range of scores exists, a matter for empirical decision, depending, on the one hand, on how many men can be individually examined and, on the other, on what margin of error in failing to uncover psychoneurotics is permissible.

In general, screening tests have either striven for *criterion-exclusive* or *criterion-inclusive* screening. In the case of criterion-exclusive screening the goal is to select a group all of whom have the desired characteristic without necessarily including within that group all who have that trait. For example, tests of vocational ability have usually assumed that there would be more applicants than openings and the function of screening tests was to select a group equal in size to the available jobs, all of whom would succeed at the job, without too much concern that other possibly successful men were excluded. On the other hand, criterion-inclusive screening strives to include within the selected group all persons who have the desired characteristic even though not all persons in the group

[3] Though paper and pencil tests, of which our test is one, are perhaps most often employed, such sensory-motor measurements as body sway and dark vision have also been successfully used. See, for example, H. J. Eysenck, "A Comparative Study of Four Screening Tests for Neurotics," *Psychological Bulletin*, Vol. 42, No. 9 (November 1945), pp. 659–662.

will have that characteristic.[4] To present this difference graphically:



Psychiatric screen tests have usually attempted inclusive screening; it has generally not been considered useful to be able to screen out a group all of whom were psychoneurotic; rather, some reasonable assurance has been wanted that all psychoneurotics would be screened from the population. From the standpoint of the practical problem with which the Research Branch was confronted—viz., to develop a psychiatric screening test to be used in induction station examination of men for the Armed Forces—this was certainly the case. The question there was not to select out a group of psychoneurotics, nor was it simply to decide how much time could be given to psychiatric evaluations of individuals and then to select the number of men who could be processed in this time in such a way as to maximize the likelihood of their being psychoneurotic. Since their goal was the detection and elimination of every disqualified psychoneurotic, psychiatric examiners at induction stations saw every man, reducing the amount of time given to each man rather than reducing the number of men seen. In order for a screening test to be useful in a situation like this, it had to aim at screening at least as many of the psychoneurotics as the brief individual examinations uncovered while reducing the number of men to be seen so that their examinations might be more thorough in the face of an acute shortage of psychiatric personnel.

Inclusive screening of this kind is adaptable only to situations in which (1) there are ways of evaluating the screened group in order to separate the men who have the screening criterion trait from the men who were falsely screened in the effort to make the screened group inclusive and (2) there can either be no cheating or where

[4] Inclusive and exclusive screening, of course, represent only the two ideal poles of a continuum. With a fine enough distribution of scores, actual screening procedures can be set at almost any point between.

cheating will simply place the malingerer in the group to be subjected to further evaluation. Psychiatric screening in the Army sought to meet these conditions: the group screened were not to be thereby rejected but simply referred for individual diagnosis, and in so far as there was a problem of malingering, it was primarily that of men trying to avoid service by claiming psychoneurotic symptoms rather than that of men trying to hide their psychiatric disturbances in order to enter the Armed Forces.[5]

As long as screening is regarded simply as a practical tool, the question of "false negatives"—men who are screened without possessing the desired trait—can be regarded thus summarily. A psychiatric screening test to be useful does not need to essay an individual prediction; except in the formal sense that it precedes the psychiatrist's evaluation and predicts his behavior,[6] detection rather than prediction is involved. If we wish to predict or to make individual diagnoses (which means, in this case, to contemplate the possibility of tests capable of replacing psychiatrists), then, of course, false negatives become as important as false positives (i.e., psychoneurotics who are not screened by the test) in indicating the sources of imperfection in the screening instrument.

Design of the Initial Study

The design of the study from which was derived the screening test, later named the Neuropsychiatric Screening Adjunct and officially adopted for use at all induction stations, was a comparison of responses of a cross section of the Army with those of a sample of psychoneurotic patients in Army hospitals to a questionnaire covering fifteen areas which the literature and psychiatric advice suggested would prove to be discriminating. This design had certain weaknesses, among which is what came to be the major question of whether the test could be transferred to the induction station situation.[7] The discussion will, however, be easier to follow if the critique of the design is reserved until the reader has a clear idea of

---

[5] Homosexuals were a possible exception to this generalization.

[6] And note that even for his behavior the prediction is only that he will find a higher proportion of the screened group to be psychiatrically disqualified than of the unscreened group, which implies that he could find the majority of the screened group acceptable without damaging the prediction.

[7] The study was not originally planned to yield such a screening test, but rather was later adapted to this purpose, a historical circumstance which serves to explain why the design was not more effectively molded to the problem. Before application at induction stations, the NSA was, however, tried out on samples of men passing through induction stations, as we shall see, in addition to the Army population.

the criterion situation in which the test was evolved and of the test itself.

The major psychiatric concern of the Army was not to exclude from service every man who was judged by psychiatrists to be psychoneurotic. Rather, psychiatric examiners were asked, in effect, to make a prediction and exclude from service only those psychoneurotics who they felt would prove to be incapable of rendering effective military service in any capacity. While the induction into and discharge from the Army of psychoneurotics in practice varied with the availability of man power, nevertheless, in policy, at least, psychoneurosis was never considered to be grounds for rejection or discharge unless it rendered a man incapable of performing "a reasonable day's work for the Army" in some assignment.[8] It was for this reason that the group of psychoneurotics currently ineffective, i.e., patients in hospitals, were chosen as the criterion group rather than using psychiatrists' diagnoses of men on duty to sort them into "normal" and psychoneurotic subgroups.[9]

This criterion of ineffectiveness left some unresolved problems, of course. At any given moment, men in the Army could be perfectly classified as effective or ineffective: either they were present for their regular duties or they were not. But it is equally obvious that at any given moment, the effective group contained some men who would later prove to be ineffective, while some of the ineffective group would return to duty. Nevertheless it was felt that a psychoneurotic breakdown under conditions of domestic Army service could at the very least be regarded as a bad prognosis for future effectiveness under possibly more trying conditions so that this group did not depart too far from the ideal concept of "ineffectiveness." And the cross section of the Army, though it contained some men who would prove to be psychiatric ineffectives, could be regarded as containing them in such negligible proportions as to approximate the polar concept of effectiveness.[10]

---

[8] See War Department Circular No. 164 (April 1944), which further provided that men with "pronounced psychiatric disorders" should not be sent overseas, but that men with other psychiatric disorders might be sent overseas and, where the disorder was "mild psychoneurosis, transient in character," could be assigned to combat duty.

[9] Ultimately, of course, a man's being a psychiatric patient in an Army hospital was also dependent on a psychiatric diagnosis. But as the Army operated, such men were not sought out, diagnosed, and hospitalized. Rather it would only be after exhibitions of gross inefficiency, constant going on sick call, etc., that these men would be culled out. They would, therefore, represent a more clear-cut group of ineffectives than a group selected by psychiatrists from among men on duty as likely to break down.

[10] Just how much and what kind of service a man should render to be considered an "effective" is itself not precisely defined. Clearly a man should complete training or

Once the criterion group was defined as those men who proved incapable of serving usefully in the Army because of psychoneurosis, there remained the problem of sampling the two universes. The cross section of the Army, to be regarded as the opposite pole, was selected in accordance with the sampling procedures usually employed by the Research Branch, but the psychoneurotic sample was more difficult. At the time of this study (January-February 1944) men who had been hospitalized for psychoneurosis in the Army could be found in six places: back at duty, with or without supplementary outpatient treatment; in special rehabilitation centers, where men received modified training and outpatient treatment; in station, regional, or general Army hospitals; or back in civilian life. However, when psychoneurotics came to the attention of medical officers in the United States and hospitalization was deemed necessary, these men were referred, as a first step, to a station hospital, that is, a hospital serving only one post or camp. After treatment and/or diagnosis here, cases were disposed of in one of the four ways suggested above; i.e., men were sent back to duty, or assigned to special centers, or referred to other hospitals, or discharged from the Army. Obviously, then, sampling at the level of station hospitals, which were the clearinghouse for all Army psychoneurotics who were hospitalized, would yield the most unselected population of hospitalized psychoneurotics from the standpoint of future disposition. While the use of a more selected sample —say, psychoneurotics to be discharged—might have more closely approximated the criterion of ineffectiveness, the station hospital population was nevertheless chosen for several reasons.

In the first place, to use a group whose final disposition was known would have introduced another variable—viz., the possibility that their responses were influenced by their knowledge that they were to get out of the Army if that group were chosen, or their presumption that this would be the outcome if men in regional or general hospitals were selected. It seemed unwise to introduce a factor

he has not been able to perform a "reasonable day's work" for the Army. But if he completes training and performs a year's work at some job before he has a psychiatric breakdown is he to be regarded as a potential ineffective who should have been excluded from the Army in the first place or did the year's service which he was able to render justify induction? At the other extreme, men who break down in combat cannot be regarded as men who should have been excluded from the Army in the first place, for we know enough about combat breakdowns to say that even among those who could not return to duty, many could have served the Army effectively in some other capacity, and many could even have gone through combat successfully if they had been subjected to shorter stretches of it.

like this. In addition, however, it is a necessary theoretical assumption in work of this kind that the psychoneurotic ineffectives be regarded as having come from among the men in the other universe, the "effectives." By sampling station hospitals and by selecting for inclusion in the sample the sixteen station hospitals which were located in and received their patients from the camps sampled to constitute the cross section of the Army, this assumption was given some basis in fact. The sample of hospitalized psychoneurotics may, then, be regarded as a representative sample of all the types of psychoneurotics hospitalized in the Army before overseas service, and as a group which came from the same universe as the cross section represented.

The Original Battery and Its Screening Results

The questionnaire basic to this study contained over a hundred personality, attitude, and background items which were grouped into fifteen areas. These fifteen areas had been selected in consultation with psychiatrists and psychiatric literature as those, on the one hand, most relevant to the psychiatric diagnosis of psychoneurosis and, on the other, likely to be susceptible to investigation through the nonintensive technique of a self-administered, anonymous questionnaire. These fifteen areas are enumerated below, with a brief description of the types of questions contained in them,[11] stated in terms of what was assumed on a priori grounds to represent good (nonpsychoneurotic) adjustment:

Aspects of Personality Development

1. *Childhood relations with parents:* Indexes the extent to which men felt that their parents had been lenient and fair and that other children in the family had not been preferentially treated.
2. *Childhood fears:* Indexes the relative absence of such childhood fears as fears of being left alone, punishment, being laughed at, etc.
3. *Childhood neurotic symptoms:* Indexes the relative absence of the traits of stuttering, stammering, nail biting, and enuresis in childhood.
4. *School adjustment:* Good adjustment is defined as having liked school and having achieved a reasonable degree of success there.
5. *Childhood fighting behavior:* Good adjustment here implies some degree of aggressiveness in boyhood as indicated by having gotten into, held one's own in, and had no fear of fights.
6. *Participation in sports:* Good adjustment is defined as having taken part in such sports as baseball, basketball, football, golf, tennis, fishing, and hunting.

[11] Part 2 of this chapter shows the exact questions used and the distribution of answers given by the cross section and by the psychoneurotic sample.

7. *Emancipation from parents:* Good adjustment here had two main components—the actual physical emancipation involved in living away from home upon occasion and the psychic emancipation implied by establishment of satisfactory heterosexual relationships.
8. *Mobility:* Well-adjusted men were expected to exhibit stability as indexed by their employment record, regularity of church attendance, and little geographical moving around.

Description of Present Personality

9. *Sociability:* Indexes a liking for and association with other people.
10. *Worrying:* Good adjustment implied a lack of worry about one's self and the future.
11. *Oversensitivity:* Good adjustment is defined as an absence of such traits as irritability, quickness to take offense, resentment of criticism, etc.
12. *Personal adjustment:* The well-adjusted man was thought to exhibit self-confidence and a lack of self-pity, depression, and anxiety.
13. *Psychosomatic complaints:* Indexes an absence of vague physical complaints like "nerves," stomach trouble, insomnia, etc.

War Motivation

14. *Identification with the war effort:* Measures the extent to which men felt that the war was worth fighting and that they had a personal stake in it.
15. *Acceptance of soldier role:* A willingness to be a soldier as judged by acceptance of fairness of being drafted, preference for job of soldier as against that of war worker, and a willingness to go overseas.

These fifteen areas had each been designed to yield a scale in the scalogram sense discussed in Chapters 1 through 9. The reasons for this design were obvious; if these areas could be shown to scale, then each of them could be represented by a single numerical variable, and moreover, the simple correlation of these scale scores with any outside variable would be precisely the multiple correlation of that variable with all of the questions entering into the scale.[12] The importance of these properties in simplifying computations is readily apparent, since, if it is desired to discriminate between two groups like normals and psychoneurotics on the basis of items which form a scale, the scores obtained by scale analysis give the best way to combine the items to get the maximum differentiation possible from the area as a whole. Accordingly, each of the areas was pretested, and all but four were found to form scales. These four—childhood fears, childhood neurotic symptoms, participation in

[12] See Chapter 5 for discussion of this point. The important and definitional property of such a scale—that from an individual's scale score can be reproduced his responses to all of the items in the scale—is not directly relevant here, though, of course, it is just because a scale approximates this condition that an outside variable correlates in the fashion described in the text.

sports, and psychosomatic complaints—did, however, form quasi scales, which lack the property of scales of reproducing the behavior of an individual on all the questions from a single numerical variable, but do possess to a large extent the property of reproducing in its zero-order correlation with an outside variable the multiple correlation to be obtained from all the items.

Therefore, each of the fifteen areas was scored with simple weights. For convenience, each of the questions was dichotomized and scored as zero or unity, unity being assigned to the responses presumed to indicate good adjustment. For some questions, there were found to be no differences between normals and psychoneurotics, but in no case did the scale or quasi scale scoring of an item prove to be the reverse of the empirically found differences on that item. This outcome indicated, first, that our a priori judgment had been correct and, second, that the scale scoring, even on a dichotomous basis, preserved the predictive efficiency of the individual items. Since these item dichotomizations and the weights assigned them had been determined on the basis of the pretest made to determine the existence of scales, rather than the main study, the possibility that the scoring was capitalizing on chance error, with the consequence that less discriminating results would be obtained in another trial, was largely eliminated.[13]

As shown in Table 1, within each of these fifteen aspects of personality, background, and adjustment, differences existed between the cross section and the psychoneurotic patients deriving from it. However, the procedure followed here is that of dichotomizing the score distribution of each variable into those scores occurring relatively more frequently among the psychoneurotics than among the Army cross section and those occurring relatively less frequently.[14] Since the point of dichotomization was determined by the score distributions in the sample under study, it is to be noted that the tendency would be to overestimate somewhat differences between psychoneurotics and nonpsychoneurotics. For this reason, we have employed here the more conservative .99 level of significance, and differences which were not found to be significant at this level are so indicated in the table.

It is clear from these data that psychosomatic complaining was

[13] It is noteworthy, however, that the scoring would have been exactly the same if the main sample had been used to determine it, an outcome which suggests the stability of scale analysis.

[14] In Chapter 9, Volume II, we have introduced the concept of *critical* and *noncritical* scores to refer to this distinction.

the area which most sharply differentiated between psychoneurotics and nonpsychoneurotics, but there remained the question of whether a combination of all the items in the battery could improve on these results or not. Accordingly, the product-moment correlations between the fifteen scores were computed and their multiple correlation with the criterion was examined.[15] From what we have said

TABLE 1

SUMMARY OF COMPARISONS BETWEEN CROSS SECTION AND PSYCHONEUROTIC PATIENTS IN FIFTEEN AREAS SCORED DICHOTOMOUSLY

(January-February 1944, Enlisted Men with No Overseas Service)

	PERCENTAGE WITH RELATIVELY LOW SCORES:*		
Area	*Cross Section*	*Neurotic patients*	*Difference*
Psychosomatic complaints	29	89	60
Personal adjustment	30	67	37
Childhood neurotic symptoms	20	53	33
Childhood fears	32	62	30
Sociability	16	45	29
Acceptance of soldier role	31	59	28
Oversensitivity	19	46	27
Worrying	25	49	24
Childhood participation in sports	40	64	24
Childhood fighting behavior	30	45	15
Childhood relations with parents	41	53	12
Identification with the war effort	24	31	7**
Mobility	49	55	6**
Emancipation from parents	36	42	6**
School adjustment	51	56	5**
Number of cases	*3,501*	*563*	

* In this table a "low" score signifies relative lack of adjustment, as this has been previously defined, in each area. In the summary scores presented in Part 2 of this chapter, these scores are shown below the dividing line The percentages given here are essentially the proportions obtained when the score distributions are dichotomized at the point which maximizes the difference between the cross section and the psychoneurotics

** Not significant at the three sigma level. All differences are significant at the two sigma level.

before, it is apparent that the multiple correlation of the criterion with these fifteen scores closely approximates the multiple correlation of the criterion on all the items in all of the fifteen areas.

The actual correlations among the fifteen variables are shown for the cross section and the psychoneurotic patients separately in

[15] Ordinarily, to predict a qualitative criterion, the discriminant function would be used, but in the special case of a dichotomous criterion the discriminant function and multiple regression are the same. Using zero and unity as numerical values for the categories of the criterion and computing product-moment correlations therefrom yields the same results as does the linear discriminant function.

TABLE 2

Intercorrelations Among Fifteen Personality and Adjustment Variables White Enlisted Men with No Overseas Service: Cross Section and Hospitalized Psychoneurotics Compared*

(January-February 1944)

	Psychosomatic complaints	*Childhood symptoms*	*Personal adjustment*	*Oversensitivity*	*Childhood fears*	*Soldier role*	*Worrying*	*Sociability*	*Participation in sports*	*Identification with war*	*Childhood fighting behavior*	*Childhood school adjustment*	*Relations with parents*	*Emancipation from parents*	*Mobility*
Psychosomatic complaints		.41 / .44	.50 / .45	.37 / .39	.32 / .35	.46 / .28	.43 / .38	.19 / .28	.16 / .06	.23 / .09	.18 / .05	.07 / .06	.09 / .16	.01 / .05	−.03 / .02
Childhood symptoms	.41 / .44		.25 / .37	.26 / .28	.24 / .33	.20 / .30	.24 / .25	.16 / .31	.16 / .21	.16 / .13	.15 / .17	.11 / .04	.10 / .15	.10 / .10	.00 / .07
Personal adjustment	.50 / .45	.25 / .37		.43 / .43	.26 / .28	.49 / .40	.41 / .44	.17 / .30	.12 / .12	.32 / .25	.08 / .07	.15 / .20	.10 / .11	.05 / .14	−.03 / .05
Oversensitivity	.37 / .39	.26 / .28	.43 / .43		.24 / .29	.22 / .17	.34 / .38	.22 / .38	.03 / .06	.25 / .22	−.10 / −.16	.17 / .22	.07 / .21	.02 / .04	.05 / .13
Childhood fears	.32 / .35	.24 / .33	.26 / .28	.24 / .29		.18 / .19	.36 / .43	.15 / .31	.08 / .17	.09 / .05	.18 / .16	.05 / .01	.02 / .12	.10 / .15	.05 / .00
Soldier role	.46 / .28	.20 / .30	.49 / .40	.22 / .17	.18 / .19		.35 / .25	.15 / .18	.16 / .25	.35 / .24	.21 / .22	.05 / .09	.07 / .02	.06 / .14	−.04 / −.05
Worrying	.43 / .38	.24 / .25	.41 / .44	.34 / .38	.36 / .43	35 / .25		.15 / .25	.08 / .07	.23 / .15	.19 / .12	.04 / .06	.02 / .11	.10 / .15	−.09 / .00
Sociability	.19 / .28	.16 / .31	.17 / .30	.22 / .38	.15 / .31	.15 / .18	.15 / .25		.18 / .21	.15 / .23	.11 / .11	.09 / .10	.07 / .17	.04 / .09	.09 / .13
Participation in sports	.16 / .06	.16 / .21	.12 / .12	.03 / .06	.08 / .17	.16 / .25	.08 / .07	.18 / .21		.08 / .13	.23 / .29	.08 / .10	.05 / .05	.12 / .13	.01 / .05
Identification with war	.23 / .09	.16 / .13	.32 / .25	.25 / .22	.09 / .05	.35 / .24	.23 / .15	.15 / .23	.08 / .13		.09 / .07	.06 / .07	.06 / .10	.08 / .05	.02 / .10
Childhood fighting behavior	.18 / .05	.15 / .17	.08 / .07	−.10 / −.16	.18 / .16	.21 / .22	.19 / .12	.11 / .11	.23 / .29	.09 / .07		.08 / −.13	−.03 / −.11	.15 / .23	−.10 / −.11
Childhood school adjustment	.07 / .06	.11 / .04	.15 / .20	.17 / .22	.05 / .01	.05 / .09	.04 / .06	.09 / .10	.08 / .10	.06 / .07	.08 / −.13		−.01 / .04	.01 / .10	.11 / .09
Relations with parents	.09 / .16	.10 / .15	.10 / .11	.07 / .21	.02 / .12	.07 / .02	.02 / .11	.07 / .17	.05 / .05	.06 / .10	−.03 / −.11	−.01 / .04		.00 / −.04	.03 / .06
Emancipation from parents	.01 / .05	.10 / .10	.05 / .14	.02 / .04	.10 / .15	.06 / .14	.10 / .15	.04 / .09	.12 / .13	.08 / .05	.15 / .23	.01 / .10	.00 / −.04		−.28 / −.27
Mobility	−.03 / .02	.00 / .07	−.03 / .05	.05 / .13	05 / .00	− 04 / − 05	−.09 / .00	.09 / .13	.01 / .05	.02 / .10	−.10 / −.11	.11 / .09	.03 / .06	−.28 / −.27	

* In each cell, the coefficient in the upper left-hand corner represents the correlation in the Army cross section; the coefficient in the lower right-hand corner, that

Table 2. Before the intercorrelations could be computed for the multiple regression, however, it was necessary to determine how to weight the cross section and the psychoneurotics in the computations. The proper equations were derived and showed that the weights should be such that the two groups in the criterion had equal variances on the prediction scores. Since the two sets of intercorrelations shown in Table 2 were so similar and since the variances of the scores did not differ much between the two groups, it did not appear worth while to solve the cumbersome equations necessary to get the exact weights, for weighting the two groups equally would approximate the equating of their variances on the prediction scores. Accordingly, the intercorrelations were recomputed for the combined population of psychoneurotics and non-psychoneurotics, regarding each of these groups as making up one

TABLE 3

INTERCORRELATION AMONG FIFTEEN PERSONALITY AND ADJUSTMENT VARIABLES AND CRITERION
WHITE ENLISTED MEN WITH NO OVERSEAS SERVICE:
CROSS SECTION AND HOSPITALIZED PSYCHONEUROTICS WEIGHTED EQUALLY AND COMBINED
(January-February 1944)

	Criterion	*Psychosomatic complaints*	*Childhood symptoms*	*Personal adjustment*	*Oversensitivity*	*Childhood fears*	*Soldier role*	*Worrying*	*Sociability*	*Participation in sports*	*Identification with war*	*Childhood fighting behavior*	*Childhood school adjustment*	*Relations with parents*	*Emancipation from parents*	*Mobility*
Criterion		.66	.38	.42	.33	.33	.35	.27	.33	.28	.12	.18	.11	.12	.08	.09
Psychosomatic complaints	.66		.55	.60	.48	.46	.50	.47	.38	.27	.19	.20	.12	.18	.07	.06
Childhood symptoms	.38	.55		.40	.36	.39	.36	.32	.34	.27	.17	.22	.12	.16	.12	.07
Personal adjustment	.42	.60	.40		.49	.37	.53	.47	.33	22	.30	.14	.19	.14	.11	.04
Oversensitivity	.33	.48	.36	.49		.36	.28	.41	.39	.13	.25	−.07	.22	.18	.05	.12
Childhood fears	33	.46	.39	.37	.36		.28	.46	.35	.21	.12	.22	.07	.12	.16	.02
Soldier role	.35	.50	.36	.53	.28	.28		.36	.25	.28	.31	.26	.10	.09	.11	−.01
Worrying	.27	.47	.32	.47	.41	.46	.36		.27	.14	.20	.19	.08	.10	.13	−.01
Sociability	.33	.38	.34	.33	.39	.35	.25	.27		.27	.21	.15	.12	.15	.06	.12
Participation in sports	.28	.27	.27	.22	.13	.21	.28	.14	.27		.12	.30	.12	.09	.14	.05
Identification with war	.12	.19	.17	.30	.25	.12	.31	.20	.21	.12		.09	.07	.10	.06	.07
Childhood fighting behavior	.18	.20	.22	.14	−.07	.22	.26	.19	.15	.30	.09		−.08	−.05	.19	−.09
Childhood school adjustment	.11	.12	.12	.19	.22	.07	.10	.08	.12	.12	.07	−.08		.03	.06	.10
Relations with parents	.12	.18	.16	.14	.18	.12	.09	.10	.15	.09	.10	−.05	.03		−.01	.06
Emancipation from parents	.08	.07	.12	.11	.05	.16	.11	.13	.06	.14	.06	.19	.06	−.01		−.28
Mobility	.09	.06	.07	.04	.12	.02	−.01	−.01	.12	.05	.07	−.09	.10	.06	−.28	

half of the combined sample.[16] These intercorrelations and the correlation of each variable with the criterion are shown in Table 3.

Rather than use the Doolittle or some similar method for computing the multiple regression, factor analysis was employed. Computing the regression exactly from Table 3 would have tended to capitalize on sampling error and would have been a somewhat

TABLE 4

LOADINGS OF COMMON FACTOR, MULTIPLE REGRESSION WEIGHTS AND RELATIVE IMPORTANCE IN TOTAL SCORE OF FIFTEEN TESTS IN A PSYCHONEUROTIC BATTERY

(January-February 1944)

		MULTIPLE REGRESSION WEIGHTS:		
Test	*Communality*	*Raw*	*Corrected**	*Number of score points possible***
Psychosomatic complaints	.61	34.5	8	120
Childhood symptoms	.41	13.0	12	48
Personal adjustment	.51	17.4	10	60
Oversensitivity	.36	10.4	4	40
Childhood fears	.34	10.1	3	57
Soldier role	.38	11.3	9	36
Worrying	.33	8.2	7	28
Sociability	.30	9.5	10	30
Participation in sports	.18	6.9	5	30
Identification with war	.13	2.8	2	10
Childhood fighting behavior	.08	4.0	4	12
Childhood school adjustment	.05	2.3	2	6
Relations with parents	.05	2.6	2	12
Emancipation from parents	.03	1.7	1	6
Mobility	.01	1.8	2	8

* The raw weights have been reduced to scale by dividing each by the standard deviation of scores on the corresponding test.

** These values are simply the corrected weight times the maximum unweighted score possible on the original test (i e , the number of items in the test)

laborious procedure as well. Instead, Thurstone's centroid method was used to extract one common factor, and it was decided that the residuals were too small to go any further. The approximate regression weights for predicting the criterion from the fifteen areas were then computed from the factor loadings by Guttman's formu-

[16] It might be pointed out that these correlations were computed only to maximize the multiple correlation with the criterion, that is, to maximize the discrimination between the cross section and the psychoneurotics. The determination of the proper weights to assign these groups was a mechanical problem which did not depend on or make any assumptions about the proportion of psychoneurotics actually in the population.

las.[17] The factor loadings and the weights derived from them were, as may be seen in Table 4, such as to suggest that the single area of psychosomatic complaints would discriminate about as well as the entire combination of fifteen areas. Random subsamples of the cross section and psychoneurotic sample were selected and scored according to these regression weights, and the result confirmed the earlier impression: when these subsamples were assigned weighted scores based on all fifteen areas, 27 per cent of the nonpsychoneurotic group and 87 per cent of the psychoneurotic group received critical scores, while the simple dichotomous scale scoring of the psychosomatic complaints area alone resulted in 29 per cent of the nonpsychoneurotics and 93 per cent of the psychoneurotics in these subsamples receiving critical scores. As a further check, the multiple regression weights of the criterion on the first eight variables in Table 3 were computed by the Doolittle method and the subsamples rescored, with exactly the same results as the use of the fifteen-area weighted scores had yielded. Clearly, there was little to be gained from preserving the other fourteen areas, and our attention was thereafter directed to attempting to improve the discrimination obtained from the psychosomatic complaints area.

Further Refinement of the Psychosomatic Complaints Score

Since the fifteen questions used in this score were known to form a quasi scale, and the criterion also fitted into this pattern, it was to be expected that different methods of weighting would not yield essentially different results. However, even a small improvement in discrimination seemed worth striving for, since a 5 per cent reduction in the number of nonpsychoneurotics receiving critical scores or a 5 per cent increase in the number of psychoneurotics screened would be of practical usefulness. The experiments with weights yielded, as we shall see, essentially negative results, although they bore out in striking fashion the stability of the quasi-scale pattern.

The results to be discussed here represent the discrimination obtained when each of three types of weights was employed:

1. *Simple dichotomous (0, 1) weights.* These were weights deriv-

---

[17] See Louis Guttman, "Multiple Rectilinear Prediction and the Resolution into Components," *Psychometrika*, Vol. 5, No. 2 (June 1940), pp. 75–99; and Louis Guttman and Jozef Cohen, "Multiple Rectilinear Prediction and the Resolution into Components: II," *Psychometrika*, Vol. 8, No. 3 (September 1943), pp. 169–183.

ing from the quasi-scale pattern of the items and consisted essentially of giving a weight of zero to those answer categories of a question which were checked by a significantly larger proportion of psychoneurotics than that in the cross section, and a weight of one to all other responses.

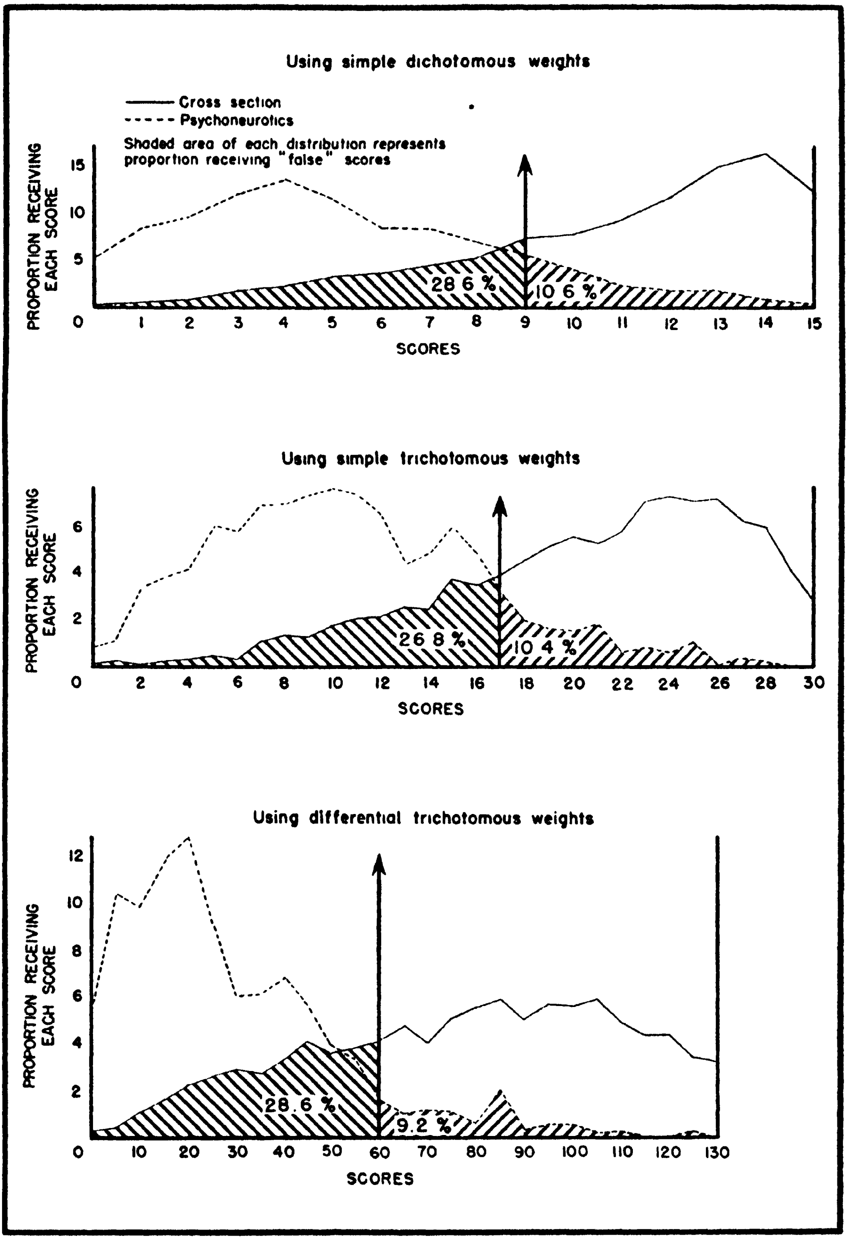
2. *Simple trichotomous (0, 1, 2) weights.* Fourteen of the fifteen items used were stated in terms of a three-point check list, while the fifteenth had four. The trichotomous weights assigned a weight of zero to the response category which occurred most disproportionately among the psychoneurotics, a weight of two to the response category at the other extreme and a weight of one to the intermediate response. An arbitrary grouping of two answer categories was made in the case of the fifteenth question in order to score it trichotomously.

3. *Differential trichotomous weights.* These were the best weights, in the sense of least squares, for discriminating between the cross section and the psychoneurotic group and were derived from the multiple regression of the criterion on the fifteen items.

Use of the simple (0, 1) weights in the area of psychosomatic complaints yielded, as we have seen before, a sharp differentiation between the cross section and the sample of psychoneurotic patients. In fact, only 28.6 per cent of the cross section received scores of 9 or less as compared with 89.4 per cent of the sample of psychoneurotics. These data are shown in the top diagram of Chart I. The increased range of scores made possible by the use of (0, 1, 2) weights resulted in a slight improvement over these results, as seen in the center diagram of Chart I. Nor were the results obtained by using weights derived from the multiple regression of the criterion on the fifteen items substantially different from the discrimination achieved through use of simple weights. (See the bottom diagram of Chart I.) In fact, in the very course of deriving these weights, it was proved mathematically that, for this set of data, the use of differential weights would not yield much improvement. This outcome was such a dramatic empirical demonstration of the scale property previously referred to—viz., that the zero-order correlation of simple scale scores with an outside variable is the multiple correlation of that variable with all of the items in the scale—that a presentation of the derivation of these weights is worth while. This appears as Part 3 of this chapter.

CHART I

DISTRIBUTION OF PSYCHOSOMATIC COMPLAINTS SCORES*
(January-February 1944)



* The position of the arrow indicates the cutting point employed to separate critical scores (on the left of the arrow) from noncritical scores. The score with which the arrow coincides is considered a critical score. Thus, in the top diagram, 0 through 9 are critical scores, 10 through 15 are noncritical.

Problems in the Application of the Test to Induction Station Screening

When it was decided to adapt the psychosomatic complaints score to induction station screening, certain questions affecting its transferability and usefulness arose. First of all, there was the easily answered question of the reliability of test scores. Then, there was a series of minor points growing out of differences in the administration of the test which would have to be introduced. The test items originally had been asked as part of a larger questionnaire through which they were scattered; did surrounding them with other questions affect the way they were responded to or would they be answered similarly when asked with no context? Moreover, the Army study employed an anonymous questionnaire; would the necessary requirement that induction station tests be signed alter men's responses?

Even if these points could be settled and dismissed there were some more serious issues to be dealt with. The test had been developed to distinguish only psychoneurotics from other men, since other categories of psychiatric disability were relatively insignificant among the men already in the Army. But, in dealing with the unselected population reporting to induction stations to be examined, psychiatrists examined men with reference to all types of psychiatric disorders. A screening test which detected psychoneurotics but permitted psychopaths and psychotics to go by undetected would be of little use, for then all men would have to be examined for these latter tendencies anyway. The question then was whether the test could be extended to psychiatric screening as well as merely to psychoneurotic screening.

Finally, there was the basic question of whether the test would really work. Thus far we had only the criterion situation in which the test was developed, a study in which the psychoneurotic and nonpsychoneurotic groups were known, and distinguishing between them was a formal exercise rather than a practical necessity. It might be, for example, that the detection of the psychoneurotics and their removal to hospitals accounted for much of the differences in their scores and the scores of men generally. If this were true, then they could not have been detected by the test so successfully prior to hospitalization. Or, it might be that the general level of psychosomatic complaining changed from civilian to military life so that the cutting point derived from the Army sample to distin-

guish critical from noncritical scores would not, in the civilian setting, furnish a sharp enough discrimination between psychiatric ineffectives and others.

The reliability of test scores. The test-retest reliability of the psychosomatic complaints score was measured by administering the test to groups of enlisted men on duty and of psychoneurotic patients and retesting them a week later. As shown in Table 5, the test scores proved to be highly reliable; in fact, about 90 per cent of each group were classified consistently, that is, their scores fell on the same side of the cutting point both times. The actual correlations of test scores with retest scores were .93 and .90 within the

TABLE 5

TEST-RETEST PSYCHOSOMATIC COMPLAINTS SCORES AMONG PSYCHONEUROTIC AND NONPSYCHONEUROTIC ENLISTED MEN (April 1944)

CLASSIFICATION OF:		PROPORTION IN EACH CLASSIFICATION AMONG:	
Test score	*Retest score*	*Enlisted men on duty*	*Psychoneurotic patients*
Noncritical	Noncritical	67.6%	21.8%
Noncritical	Critical	3.0	4.0
Critical	Noncritical	6.4	6.9
Critical	Critical	23.0	67.3
		100.0%	100.0%
Number		*299*	*101*

nonpsychoneurotic and the psychoneurotic groups respectively. Indeed these relationships were so high that it was feared that they might be in part a result of carryover from the test to the retest situation, especially in view of the small period of time between them. But estimates of the lower limit of the reliability coefficient, using the Kuder-Richardson formulas, were comparably high, being .87 and .82 respectively.

The effect of context. In order to check on the possible influence on test scores which asking the items as part of a larger questionnaire in which they were embedded might have had, a form was prepared which placed the test items at the beginning of the questionnaire so that the answers to them could not be influenced by preceding questions. This new form was administered in conjunction with the original form to a small group of enlisted men on duty and another group of enlisted men who were psychoneurotic pa-

tients, a random half of each group filling out each form. As shown in Table 6, there were no significant differences attributable to the type of form used.

The effect of anonymity. There was some question of whether signing their names would make men more or less cautious in claiming psychosomatic symptoms than they were under strictly anonymous conditions. On the one hand it was argued that identification and the consequent possibility of checking up on their responses would make men more cautious in claiming symptoms. On the other hand, identification and the implication that test scores would have some bearing on one's draft classification or disposition in the

TABLE 6

RELATION OF PRESENCE OR ABSENCE OF CONTEXT TO PSYCHOSOMATIC COMPLAINTS SCORES AMONG ENLISTED MEN ON DUTY AND PSYCHONEUROTIC PATIENTS

(April 1944)

	PROPORTIONS RECEIVING CRITICAL SCORES:	
Type of form answered	*Nonpsychoneurotics*	*Psychoneurotics*
With context	33.8% (*145*)	74.0% (*50*)
Without context	25.3 (*154*)	74.5 (*51*)
Difference	8.5	−0.5
Critical ratio of difference	*1.60*	*0.02*

Army might encourage some men to exaggerate their symptoms. Obviously, if the use of an identified form made a difference, then the test standardized under another set of conditions would not necessarily perform adequately.

In three studies, two made in the Army and one in an induction station, small but consistent differences were found between men answering the test anonymously and men asked to sign their names to their test forms. In each case, critical scores occurred somewhat less often among the identified men. The data from the induction station study in which random halves of a group of men being examined there filled out signed and anonymous forms, with no explanation of the purpose of the schedules, are shown in Table 7. These differences are not in themselves statistically significant, nor do they constitute sufficient replications to confirm the existence of a real tendency. But even if other studies would have confirmed the difference, it was of such a small order as to be of no practical importance.

The effect of change of situation and criterion. Up to this point, when we have spoken of distinguishing psychoneurotics from non-psychoneurotics our data have been derived from a comparison of men known or presumed to be psychoneurotic ineffectives (the criterion group) with a group known or presumed to be free of the criterion. This is, of course, a standard procedure in developing tests of this kind, but the crucial question on which such a test's

TABLE 7

Distribution of Psychosomatic Complaints Scores among Anonymous and Identified Preinduction Examinees (March 1944)

Score	Proportions receiving each score: *Anonymous men*	Proportions receiving each score: *Identified men*	Proportions receiving given score or lower: *Anonymous men*	Proportions receiving given score or lower: *Identified men*
30	0.4%	2.5%	100.0%	100.0%
29	3.4	3.8	99.6	97.5
28	2.9	8.8	96.2	93.7
27	7.1	5.9	93.3	84.9
26	5.4	7.1	86.2	79.0
25	4 6	6.3	80.8	71.9
24	6.3	8.4	76.2	65.6
23	8.3	6.7	69.9	57.2
22	8.8	5.9	61.6	50.5
21	7.5	5.5	52.8	44.6
20	3.8	3.8	45.3	39.1
19	7.1	3.8	41.5	35.3
18	3.8	3.8	34.4	31.5
17	5.4	4.6	30.6	27.7
16	2.5	3.4	25.2	23.1
15	2.9	4.2	22.7	19.7
14	4.2	3.4	19.8	15.5
13	2.9	2.1	15.6	12.1
12	3.4	2.9	12.7	10.0
11	1.7	1.3	9.3	7.1
10	1.7	0.4	7.6	5.8
9	1.7	1.7	5.9	5.4
8	2.1	0.4	4.2	3.7
7	1.3	0.8	2.1	3.3
6	0.0	1.7	0.8	2.5
5	0.0	0.4	0.8	0.8
4	0.0	0.0	0.8	0.4
3	0.0	0.0	0.8	0.4
2	0.0	0.4	0.8	0.4
1	0.8	0.0	0.8	0.0
0	0.0	0.0	0.0	0.0
Total	100.0%	100.0%		
Number	*238*	*239*		

usefulness depends is whether it could have achieved similar results when applied to an unselected population prior to its separation into the two components in which we are interested; viz., psychoneurotics and nonpsychoneurotics. That is to say, for screening purposes our results must be translatable into the statement that if (roughly) the third of an unselected group of enlisted men who scored lowest on our test were to be examined, there would be found among this third of the group close to 90 per cent of all those in the entire group who would later become psychoneurotic ineffectives. This in turn implies that the test would lend itself to a practical screening procedure only if it could be shown that the scores obtained by psychoneurotic ineffectives were not appreciably affected either by their hospitalization or by developments between their induction and their hospitalization; in other words, that as large a percentage of them would have received critical scores while still in an ordinary milieu as did in the hospital setting. It was also of some relevance to the test's practical usefulness to know how stable the scores of nonpsychoneurotics remained. If there were any appreciable shifts in their scores from civilian to military life, then the test might require restandardization in terms of the scores to be considered critical.

Ideally, these crucial points should have been tested by a study in which a group of men being examined for induction were given the psychoneurotic test and then followed up after a reasonable length of time in their Army careers, say a year or two. It could then have been determined what had become of these men, and their earlier scores classified according to whether they had become psychoneurotic ineffectives or not. A comparison of the score distributions for the "failures" and "successes" would then have clearly indicated how effectively the test actually discriminated before the fact. Moreover a retesting of these men would have permitted a determination of the direction and extent of score shifting accompanying the change from civilian to military life. At the time the NSA was being developed, however, the need for it or a similar test seemed urgent, and there simply was not time to undertake this definitive check.[18]

As an alternative to the procedure just outlined, there was the

[18] Such a study was later undertaken. A sample of about 5,000 men inducted into the Army at approximately the same time in different parts of the country was gathered and the preinduction NSA scores of the men were secured. The ending of the war rather soon afterwards, however, made the follow-up impractical.

faster though less exact technique of comparing the scores obtained in the samples of enlisted men studied with those of comparable groups of men prior to induction. But in order for comparability even to be approximated, it was necessary that the men tested prior to induction be separable into the criterion group of psychoneurotic ineffectives and the contrast group of effectives. Yet the criterion was defined in terms of actually proving to be ineffective in the Army situation, a reaction which could not be determined prior to induction and service. Psychiatrists' diagnoses were clearly the only substitute way of separating the criterion and contrast groups, but their use meant substituting for the original criterion of *actually* becoming ineffective, psychiatrists' *predictions* that certain men would prove to be psychoneurotic ineffectives if they were inducted. The two criteria would have been interchangeable only if it could be assumed that the diagnostic decisions psychiatrists made on the basis of brief induction station examinations were completely accurate in terms of the original criterion, that is, that every man found by psychiatrists to be acceptable for service remained effective and that every man they rejected would have become ineffective if inducted. But the very fact that there were breakdowns and discharges of men once found acceptable indicated the less than perfect correspondence of the two criteria. In fact, one study made by the Surgeon General's Office showed that the men inducted from an induction station where a high rate of psychiatric rejections prevailed (presumably, therefore, a rather selected population psychiatrically) later had as high a rate of discharges for psychiatric defects as a presumably less selected group of men inducted from a neighboring station which had a low rate of psychiatric rejections. This study, as well as others showing a rather high survival rate in the Army for groups of men who would ordinarily have been rejected by psychiatrists, suggested that there might even be relatively little relationship between the two criteria.[19]

On the other hand, though the Army's goal was to increase efficient use of man power both by reducing rejections of men who would have succeeded in the Army and by decreasing the inductions

[19] It should be remembered that it was far easier to decide whether or not a man should be diagnosed psychoneurotic than it was to decide whether he should be rejected for this reason, especially when not too much was known about the psychological factors making for ineffectiveness or survival in the Army. Many psychiatrists therefore tended to regard any psychoneurotic as a bad risk for the Army. Others tended to assume that psychoneurotics who had developed minimally adequate adjustments to their civilian situation could also do so in military life, and so to recommend their induction. Thus, errors of judgment could be made on both sides.

of men who would become ineffective, it is certain that any test which did not reproduce rather well the psychiatrists' induction station judgments would have been impossible of adoption, even if it did discriminate well with reference to the criterion of ineffectiveness which the psychiatrists were predicting. The reason for this is obvious: it must be recalled that the test scores themselves did not lead directly to disposition; rather, they peeled off a fairly sizable group, among whose number the psychoneurotics to be rejected were to be discovered by psychiatric examination. The standards of judgment which psychiatrists would use in examining this subgroup would naturally be those they employed prior to the introduction of the test. If the diagnostic criteria implicit in the test were not rather close to those ordinarily employed by psychiatrists, then they would not in fact find a disproportionate number of psychoneurotics in the screened subgroup and would, therefore, have no confidence in the test. And, even if a screening test simply reproduced the behavior of induction station psychiatrists without in any way improving on their efficiency in economizing man power, it at least had the virtue of saving psychiatrists' time, which was itself critically scarce.

At any rate, a small trial of the test was undertaken at one induction station and certain conclusions were drawn from it. In this study, Research Branch personnel administered the test to a random group of men being processed and then removed the tests from the station. Psychiatrists examined the men in the usual fashion, without knowledge of either the men's test scores or even the items in the test itself. These independent diagnoses were then collected and the men's test scores were compared with them. The basic data from this trial are shown in Table 8.

As we see here, when critical scores were divided from noncritical scores at the same point as was arrived at by analysis of the Army data, then 18.0 per cent of the men found acceptable for service received critical scores as compared with 81.8 per cent of the men rejected for psychiatric reasons. It will be recalled that, in the Army study, 26.8 per cent of the cross section and 89.6 per cent of the hospitalized psychoneurotic patients had received critical scores. At first glance, then, the test was about as successful in screening with reference to the original criterion.

In comparison with the induction station data, however, the Army data seemed to indicate that there was a general downward shifting of scores in the Army milieu. That is to say, men already

TABLE 8

Psychosomatic Complaints Scores of Candidates for Induction, Classified by Disposition: One Induction Station
(March 1944)

	Proportions receiving each score				Proportions receiving given score or less			
Score	*All men*	*Men acceptable for induction*	*Men rejected on physical grounds*	*Men rejected on psychiatric grounds**	*All men*	*Men acceptable for induction*	*Men rejected on physical grounds*	*Men rejected on psychiatric grounds**
30	2.5%	3.7%	0.0%	0.0%	100.0%	100.0%	100.0%	100.0%
29	3.8	4.8	2.9	0.0	97.5	96.3	100.0	100.0
28	8.8	9.7	14.7	0.0	93.7	91.5	97.1	100.0
27	5 9	7.8	3.0	0.0	84.9	81.8	82.4	100.0
26	7.1	7.2	11.8	2.6	79.0	74.0	79.4	100.0
25	6.3	7.2	8.8	0.0	71.9	66.8	67.6	97.4
24	8.4	9.1	11.8	2.6	65.6	59.6	58.8	97.4
23	6.7	9.1	0.0	2.6	57.2	50.5	47.0	94.8
22	5.9	6.0	8.8	2.6	50.5	41.4	47.0	92.2
21	5.5	4.8	14.7	0.0	44.6	35.4	38.2	89.6
20	3.8	4.8	0.0	2.6	39.1	30.6	23.5	89.6
19	3.8	4.2	3.0	2.6	35.3	25.8	23.5	87.0
18	3 8	3.6	5.9	2.6	31.5	21.6	20.5	84.4
17	4.6	5.4	0.0	5.3	27.7	18.0	14.6	81.8
16	3.4	3.0	2.9	5.3	23.1	12.6	14.6	76.5
15	4.2	3.0	2.9	10.6	19.7	9.6	11.7	71.2
14	3.4	1.8	3.0	10.6	15.5	6.6	8.8	60.6
13	2.1	1.8	0.0	5.3	12.1	4.8	5.8	50.0
12	2.9	0.6	2.9	13.2	10.0	3.0	5.8	44.7
11	1.3	0.0	2.9	5.3	7.1	2.4	2.9	31.5
10	0.4	0.0	0.0	2.6	5.8	2.4	0.0	26.2
9	1.7	0.6	0.0	7.9	5.4	2.4	0.0	23.6
8	0.4	0.6	0.0	0.0	3.7	1.8	0.0	15.7
7	0 8	0.6	0.0	2.6	3.3	1.2	0.0	15.7
6	1 7	0 6	0.0	7.9	2.5	0.6	0.0	13.1
5	0.4	0.0	0.0	2.6	0.8	0.0	0.0	5.2
4	0.0	0.0	0.0	0.0	0.4	0.0	0.0	2.6
3	0.0	0.0	0 0	0.0	0.4	0.0	0.0	2.6
2	0.4	0.0	0.0	2.6	0.4	0.0	0.0	2.6
1	0.0	0.0	0 0	0.0	0.0	0.0	0.0	0.0
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Total	100.0%	100.0%	100.0%	100.0%				
Number of cases	*238*	*166*	*34*	*38*				

* Includes men rejectable for both physical and psychiatric reasons.

in the Army scored less favorably on the average than men acceptable for service but not yet in the Armed Forces. The difference between the proportions receiving critical scores was the more remarkable because the men already in the Army had been subjected to more psychiatric selection than the potential inductees, through the breakdown and hospitalization of some of its original numbers. In so far as the military patients could be assumed to have come from among the small proportion of men who received critical scores but were inducted,[20] we would expect that the Army group would appear to have better scores than the induction station group, even if there had been complete stability of individual scores. Instead we found a difference in the other direction.

Taken by themselves, these data would appear to suggest that an upward revision of the highest score regarded as critical at induction stations was in order. However, these data were derived from only one induction station, and there was no way of knowing how representative they would be of other stations in other parts of the country.[21] Moreover, though the percentage of critical scores among men acceptable for induction appeared lower than the proportion among men already in the Army, as far as could be told from the small number of cases available, the same was true of the proportion of psychiatric rejections screened at induction stations in comparison with the figure among psychoneurotics hospitalized in the Army. Yet, despite the generally upward shift of score distributions for both acceptable men and unacceptable men at induction stations in comparison with the Army data, these two distributions intersected at the same scale score for the induction station data as they had in the Army data. The point of intersection is, of course, the most efficient cutting point to use in separating critical from noncritical scores, though it need not be used if some other point better implements the objective of the test.

In view of the coincidence of the cutting points in the two studies, it was decided that for the time being, at least, no revision of cutting point should be made. This reasoning was supported by at least two further considerations. First, since the data for psychiatric rejections consisted of only thirty-eight cases, it was felt that the

[20] This is, of course, the crucial and untested question which makes it impossible to say definitively whether the test screened well in practice with reference to the criterion of future ineffectiveness.

[21] Evidence gathered later confirmed the reality of the apparent shift in scores, but this information was not available during the developmental phases of the test.

general shape of the distribution of these scores, as reflected in the point of intersection, was more reliable than the exact percentage distribution. That is to say, no matter where the cutting point had been set, the percentage of psychiatric rejections being screened at that point could not be determined with any reliability, so it was not very useful to attempt to raise the cutting point to the level at which some particular percentage was screened. Second, the results obtained were based on the application of a test for psychoneurotics to all psychiatric rejections. After extensive discussion with psychiatrists, the conclusion was reached that questions aimed more explicitly at the detection of psychotics and psychopaths would have to be added. Since this revision of the test would change the score distributions anyway, there was little to be gained from fixing on any particular cutting point until the new distributions were available, so the original cutting point was continued.

Questions added to detect "critical signs." The questions finally added in the interest of extending the test's psychiatric applicability were eight in number.[22] These were not combined with the original fifteen items which continued to be scored as we have described, but were treated as qualitative. Possession of a *critical sign,* as an affirmative response to any of these questions was designated, was considered sufficient reason for referral to the psychiatrist for evaluation, regardless of whether the numerical score was critical or not. Thus, the final test score consisted of two parts: the numerical score and a set of symbols to indicate what critical signs had been claimed by the respondent. Men with scores critical in either respect (i.e., a low numerical score or the possession of one or more critical signs) were to be considered screened by the test.

In this form the test was named the Neuropsychiatric Screening Adjunct (NSA).[23] It was officially adopted for use at all induction

---

[22] The exact questions, to be answered by a simple "yes" or "no," were:

1. Have you ever had stomach ulcers?
2. Do you ever take dope?
3. Have you ever had fits or convulsions since you were ten years old?
4. Did you ever have a nervous breakdown?
5. Were you ever a patient in a mental hospital (because of your nerves)?
6. Were you ever sent to reform school?
7. Have you ever gotten into serious trouble or lost your job because of drinking?
8. Do you ever wet the bed? (This means urinate in bed, *not* wet dreams.)

[23] The naming of the test was a historical accident, which need not concern us here, except to point out that the test obviously had no bearing on the detection of the neurologically unfit, who may, for our purposes, be regarded as a subcategory of the physical rejections.

stations beginning in October 1944.[24] The procedures outlined for its use called for its administration to all men passing the minimum literacy test, and the psychiatric examination of those who were screened by it, i.e., received critical scores and/or critical signs. The psychiatric examination of men who took the test and were not screened by it was left discretionary, but it was intended that until its use had been routinized and a further check of its operation could be made, all men be seen by the psychiatrists regardless of test scores. As it turned out, the extreme need for the NSA to relieve the pressure on psychiatrists had diminished by the time of its introduction, and it was never considered necessary to make full use of the test to pass certain men without individual examinations.

PART 2

QUESTIONS USED IN INITIAL SURVEY AND THEIR DISCRIMINATING UTILITY

Comparisons are shown here of responses of 3,501 white enlisted men with no overseas service and 563 psychoneurotic patients in Army hospitals, in January-February 1944.[25] "No answers" are omitted.

The attitude data are, of course, men's subjective reports and may not be accurate measures of the actual incidence of the behaviors the questions refer to.

The items cover the fifteen areas described earlier in this chapter.

1. BACKGROUND

How old were you on your last birthday?

	Neurotic patients	*Cross section*
30 and over	40%	22%
25–29	27	25
—————————*		
21–24	22	35
Under 21	11	18
	100%	100%

[24] See War Department Memorandum No. 40–44 (19 September 44) "Psychological Examining for Neuropsychiatric Tendencies."

[25] S–99.

* This line marks the cutting point at which questions were divided for the purpose of setting up dichotomized series.

How far did you go in school?

	Neurotic patients	*Cross section*
College	9%	16%
High school graduate	15	32
Some high school	37	31
Grade school	39	21
	100%	100%

Are you single, married, divorced, separated, or widowed?

	Neurotic patients	*Cross section*
Single	43%	61%
Married	52	36
Divorced, separated, or widowed	5	3
	100%	100%

Where did you live most of the time before you were sixteen years old?

	Neurotic patients	*Cross section*
Farm	29%	26%
Small town (under 2,500)	11	13
Town (2,500–25,000)	14	18
City (25,000–100,000)	15	15
Large city (over 100,000)	31	28
	100%	100%

How long have you been in the Army?

	Neurotic patients	*Cross section*
Over 2 years	13%	20%
Over 1 year to 2 years	41	49
Over 6 months to 1 year	19	19
6 months or less	27	12
	100%	100%

How many brothers and sisters *older* than you do you have? How many brothers and sisters *younger* than you do you have?

	Neurotic patients	*Cross section*
Only child	6%	8%
Youngest child	22	21
Intermediate child	52	49
Oldest child	20	22
	100%	100%

Did your parents always live together up to the time you were sixteen years old?

	Neurotic patients	*Cross section*
Yes	71%	77%
No		
Death	17	15
Divorce or separation	10	6
Other	2	2
	100%	100%

So far as you know, has anyone in your family ever had a nervous breakdown? (Do *not* include yourself if you think you have had a nervous breakdown.)

	Neurotic patients	*Cross section*
No	25%	60%
Don't know	29	18
Yes	46	22
	100%	100%

. CHILDHOOD RELATIONS WITH PARENTS

Summary score for area

A high score represents a feeling that parents were lenient and fair, that other children in the family were not preferentially treated, and a low score represents the opposite.

Score	*Neurotic patients*	*Cross section*
6, 5, 4	16%	22%
3	31	37
2	30	26
1, 0	23	15
	100%	100%

When you were growing up, who would you say was your *mother's* favorite child?

	Neurotic patients	*Cross section*
Only child	8%	8%
I was	15	17
There were no favorites	54	56
An older child	10	9
A younger child	13	10
	100%	100%

When you were growing up, who would you say was your *father's* favorite child?

	Neurotic patients	*Cross section*
Only child	8%	8%
I was	11	11
There were no favorites	58	58
An older child	11	11
A younger child	12	12
	100%	100%

Do you think that your parents (or the people who brought you up) were stricter than other parents or not as strict as other parents?

	Neurotic patients	*Cross section*
Not as strict	15%	15%
About as strict	67	69
Stricter	18	16
	100%	100%

When you were growing up, about how often were you punished by your parents (or the people who brought you up)?

	Neurotic patients	*Cross section*
Not often enough	17%	23%
About the right number of times	72	74
Too often	11	3
	100%	100%

When your parents (or the people who brought you up) punished you, did you usually deserve it or not?

	Neurotic patients	*Cross section*
Always deserved it	50%	64%
Deserved it most of the time	36	29
Deserved it about half the time or less	14	7
	100%	100%

When you had actually done something wrong and were punished for it, did you usually get an easier or a harder punishment than other kids you knew?

	Neurotic patients	*Cross section*
Easier	12%	14%
About the same	71	73
Harder	17	13
	100%	100%

3. CHILDHOOD FEARS

Summary score for area

In the area of childhood fears, a high score represents fewer fears reported in childhood of such things as snakes, punishment, being on high places, etc.

Score	*Neurotic patients*	*Cross section*
19 (no fears)	9%	25%
18–16 (1–3 fears)	29	43
15–12 (4–7 fears)	31	25
11–0 (8 or more fears)	31	7
	100%	100%

When you were a kid, how afraid were you of:

	Neurotic patients	*Cross section*
Strangers		
Not at all	54%	66%
A little	32	30
Very much	14	4
	100%	100%
Being with girls		
Not at all	53%	67%
A little	33	26
Very much	14	
	100%	100%
Thunderstorms		
Not at all	44%	60%
A little	36	31
Very much	20	9
	100%	100%
Being left alone		
Not at all	41%	48%
A little	35	43
Very much	24	9
	100%	100%

	Neurotic patients	Cross section
Large animals		
Not at all	43%	52%
A little	35	37
Very much	22	11
	100%	100%
Being shut up in a room or closet		
Not at all	46%	62%
A little	29	27
Very much	25	11
	100%	100%
Falling		
Not at all	32%	47%
A little	35	42
Very much	33	11
	100%	100%
Walking by a graveyard at night		
Not at all	38%	51%
A little	35	36
Very much	27	13
	100%	100%
The Devil		
Not at all	56%	65%
A little	20	20
Very much	24	15
	100%	100%
Thoughts of death		
Not at all	45%	52%
A little	28	33
Very much	27	15
	100%	100%
Being on high places		
Not at all	34%	49%
A little	27	36
Very much	39	15
	100%	100%

	Neurotic patients	*Cross section*
Family quarrels		
Not at all	39%	49%
A little	33	34
Very much	28	17
	100%	100%
Being punished		
Not at all	27%	32%
A little	41	51
Very much	32	17
	100%	100%
Being laughed at by other boys		
Not at all	36%	46%
A little	29	36
Very much	35	18
	100%	100%
Sharp knives		
Not at all	38%	52%
A little	32	29
Very much	30	19
	100%	100%
Getting bawled out		
Not at all	22%	32%
A little	39	48
Very much	39	20
	100%	100%
Being called on to recite in class		
Not at all	27%	36%
A little	34	42
Very much	39	22
	100%	100%
Getting a bad report card from school		
Not at all	32%	38%
A little	37	38
Very much	31	24
	100%	100%

	Neurotic patients	Cross section
Snakes		
Not at all	23%	26%
A little	24	35
Very much	53	39
	100%	100%

4. CHILDHOOD NEUROTIC SYMPTOMS

Summary score for area

The area of childhood neurotic symptoms is concerned with such things as general health, stuttering, stammering, nail biting and enuresis. A high score represents absence of these symptoms.

Score	*Neurotic patients*	*Cross section*
4 (no symptoms)	18%	46%
3 (1 symptom)	29	34
2 (2 symptoms)	28	15
1, 0 (3 or 4 symptoms)	25	5
	100%	100%

As far as you know, were you a healthy child or a rather sickly one?

	Neurotic patients	*Cross section*
Very healthy	21%	51%
Fairly healthy	40	36
Don't know	7	3
Rather sickly	32	10
	100%	100%

Did you ever bite your fingernails when you were a child?

	Neurotic patients	*Cross section*
Never	40%	55%
Sometimes	29	28
Often	31	17
	100%	100%

When you were growing up, did you have any trouble with stuttering or stammering in your speech?

	Neurotic patients	*Cross section*
Never	50%	74%
Sometimes	35	20
Often	15	6
	100%	100%

How old were you when you stopped wetting the bed?

	Neurotic patients	*Cross section*
Less than 4 years old	56%	76%
4–9 years	22	16
10–16 years	13	6
After 16 years	4	1
Still wet the bed	5	1
	100%	100%

5. CHILDHOOD SCHOOL ADJUSTMENT

Summary score for area

A high score indicates a liking for school and a reasonable degree of success there.

Score	*Neurotic patients*	*Cross section*
3	10%	19%
2	34	30
1	29	32
0	27	19
	100%	100%

As a child how did you feel about going to school?

	Neurotic patients	*Cross section*
Really wanted very much to go	46%	49%
Didn't care very much whether I went or not	45	45
Did not at all want to go	9	6
	100%	100%

When you were a kid, how often would you say you played hookey from school (stayed away as much as a day)?

	Neurotic patients	*Cross section*
Never	27%	28%
Only once or twice	33	40
Several times	31	27
Very often	9	5
	100%	100%

What kind of grades did you usually get when you were in school?

	Neurotic patients	*Cross section*
Very high	4%	5%
High	17	27
Medium	62	64
Low	11	3
Very low	6	1
	100%	100%

6. CHILDHOOD FIGHTING BEHAVIOR

Summary score for area

Good adjustment is defined as some degree of aggressiveness as indicated by getting into, holding one's own in, and not being afraid of fights.

Score	*Neurotic patients*	*Cross section*
3	29%	43%
2	26	27
1	23	19
0	22	11
	100%	100%

How often did you get into fights when you were a kid?

	Neurotic patients	*Cross section*
Very often	21%	16%
Sometimes	38	46
Almost never	41	38
	100%	100%

How did you feel about fighting as a kid?

	Neurotic patients	*Cross section*
Really liked fighting	6%	12%
Didn't particularly like or dislike it	38	54
Didn't like fighting at all	56	34
	100%	100%

Were you a good fighter as a kid?

	Neurotic patients	*Cross section*
Yes	22%	39%
Undecided	36	33
No	42	28
	100%	100%

7. CHILDHOOD PARTICIPATION IN SPORTS

Summary scores for area

(a) Number of sports engaged in

	Neurotic patients	*Cross section*
6, 5, 4	6%	12%
3, 2	30	48
1	26	25
0	38	15
	100%	100%

(b) Type of sports engaged in

	Neurotic patients	*Cross section*
Both body-contact and non-body-contact sports	26%	49%
Body-contact sports only	14	19
Non-body-contact sports only	22	17
None	38	15
	100%	100%

As a boy (before you were 18 years old), how much did you take part in these sports?

	Neurotic patients	*Cross section*
Baseball		
A lot	31%	54%
A little	43	36
Not at all	26	10
	100%	100%

	Neurotic patients	Cross section
Fishing or hunting		
A lot	39%	50%
A little	29	31
Not at all	32	19
	100%	100%
Football		
A lot	18%	38%
A little	30	37
Not at all	52	25
	100%	100%
Basketball		
A lot	18%	32%
A little	32	42
Not at all	50	26
	100%	100%
Tennis		
A lot	7%	12%
A little	20	32
Not at all	73	56
	100%	100%
Golf		
A lot	4%	7%
A little	15	22
Not at all	81	71
	100%	100%

8. EMANCIPATION FROM PARENTS

Summary score for area

This score has two main components: the actual physical emancipation involved in leaving home and the psychic emancipation implied by establishment of satisfactory heterosexual relationships; so a high score in this area represents good adjustment in both respects.

Score	*Neurotic patients*	*Cross section*
6, 5	13%	14%
4, 3	45	50
2, 1	29	29
0	13	7
	100%	100%

How old were you when you started living away from your parents or the family that brought you up?

	Neurotic patients	*Cross section*
Less than 15 years old	10%	6%
15–17 years	13	11
18–21 years	22	29
22 years and over	18	14
Never lived away	37	40
	100%	100%

Before you came into the Army, what was the longest you had ever been away from your family or *anybody related to you?* (By "family" is meant your parents, brothers and sisters, or people who brought you up.)

	Neurotic patients	*Cross section*
1 year or more	27%	26%
1 month to 1 year	29	34
1 week to 1 month	19	21
Less than 1 week	15	12
Never away	10	7
	100%	100%

Before you came into the Army, what was the longest you had ever been away *at any one time* from your parents or the family that brought you up? (If you left home to get married, count the time you were away after you were married.)

	Neurotic patients	*Cross section*
1 year or more	35%	31%
1 month to 1 year	26	34
1 week to 1 month	18	21
Less than a week	15	9
Never away	6	5
	100%	100%

How old were you when you first started going around by yourself on dates with girls (*not* on double dates or in large groups)?

	Neurotic patients	*Cross section*
Less than 15 years old	7%	7%
15–16 years	20	30
17–18 years	33	38
19–20 years	20	15
21 years or over	12	6
Never dated	8	4
	100%	100%

Did you usually have dates with girls more often or less often than most other fellows of about your own age that you knew?

	Neurotic patients	*Cross section*
More often	13%	13%
About the same	45	61
Not as often	42	26
	100%	100%

How old were you when you first started going steady with one girl?

	Neurotic patients	*Cross section*
Less than 18 years old	16%	20%
18–20 years	35	40
21–23 years	19	16
23 years or more	15	8
Never went steady	15	16
	100%	100%

9. MOBILITY

Summary score for area

A high score represents a stable employment record, regular church attendance and little geographical moving around.

Score	*Neurotic patients*	*Cross section*
4	13%	14%
3	32	37
2	35	36
1, 0	20	13
	100%	100%

Before you came into the Army, how many different states of the United States had you been in to live or work?

	Neurotic patients	*Cross section*
One	45%	43%
Two	23	21
Three	11	13
Four	7	8
Five or more	14	15
	100%	100%

How many jobs did you have in the last two years before you came into the Army?

	Neurotic patients	*Cross section*
One	45%	47%
Two	29	32
Three	10	12
Four or more	12	6
Never had a job	4	3
	100%	100%

Before you came into the Army, had you ever been fired or asked to resign from a job?

	Neurotic patients	*Cross section*
No	78%	89%
Never had a job	4	3
Yes	18	8
	100%	100%

Before you came into the Army, about how often did you go to church?

	Neurotic patients	*Cross section*
Once a week or more	34%	39%
One to three times a month	26	29
Several times a year	22	16
Almost never	18	16
	100%	100%

10. SOCIABILITY

Summary score for area

A high score represents a liking for and association with other people.

Score	*Neurotic patients*	*Cross section*
3	22%	43%
2	33	41
1	26	12
0	19	4
	100%	100%

How would you say the people you know feel about you?

	Neurotic patients	*Cross section*
Almost all of them like me	46%	53%
Most of them like me	39	43
A few of them like me	12	3
Almost none of them like me	3	1
	100%	100%

Before you came into the Army did you usually go around with a bunch of others or by yourself?

	Neurotic patients	*Cross section*
With a bunch of others	18%	38%
With just one or two others	49	52
By myself	33	10
	100%	100%

On the whole, before you came into the Army did you usually like to be by yourself or to be with other people?

	Neurotic patients	*Cross section*
With other people	44%	78%
Undecided	15	7
By myself	41	15
	100%	100%

11. IDENTIFICATION WITH THE WAR EFFORT

Summary score for area

Identification with the war effort is defined as a feeling that the war is worth fighting and a recognition of one's own personal stake in it, with a high score representing identification and a low score, lack of identification.

Score	*Neurotic patients*	*Cross section*
5	43%	49%
4	26	27
3, 2	20	18
1, 0	11	6
	100%	100%

Do you think you have as much of a personal stake in this war as anybody else?

	Neurotic patients	*Cross section*
Yes	68%	77%
I think so but I'm not sure	13	9
Undecided	12	7
No	7	7
	100%	100%

In your opinion is the United States fighting for things that *you* feel are worth fighting for?

	Neurotic patients	*Cross section*
Yes	84%	89%
I think so, but I'm not sure	6	5
Undecided	6	4
No	4	2
	100%	100%

Do you think it would make much difference to *you personally* if we did not win this war?

	Neurotic patients	*Cross section*
Would make a great deal of difference	85%	94%
Quite a bit	10	5
Not so much	4	1
Not any	1	0
	100%	100%

Do you ever get the feeling that this war is not worth fighting?

	Neurotic patients	*Cross section*
Never	64%	65%
Only once in a great while	17	18
Sometimes	13	12
Very often	6	5
	100%	100%

Do you think the things we are fighting for are worth risking your life for?

	Neurotic patients	*Cross section*
Yes	74%	83%
I think so, but I'm not sure	13	9
Undecided	8	5
No	5	3
	100%	100%

12. ACCEPTANCE OF SOLDIER ROLE

Summary score for area

A high score in the area of acceptance of soldier role represents a feeling that being drafted was fair in one's own case, a choice of the job of a soldier in preference to one as a war worker, and a desire for overseas service.

Score	*Neurotic patients*	*Cross section*
4, 3	16%	46%
2	25	23
1	32	20
0	27	11
	100%	100%

At the time you came into the Army did you think you should have been deferred?

	Neurotic patients	*Cross section*
I wasn't drafted	18%	25%
No	38	50
Yes		
Health	35	9
Dependents	5	8
Job	3	6
Other	1	2
	100%	100%

If it were up to you to choose, do you think you could do more for your country as a soldier or as a worker in a war job?

	Neurotic patients	*Cross section*
Soldier	10%	40%
Undecided	18	15
War worker	72	45
	100%	100%

If it were up to you and you yourself *had to decide*, would you choose to be a soldier or a civilian?

	Neurotic patients	*Cross section*
Soldier	21%	44%
Undecided	23	13
Civilian	56	43
	100%	100%

If it were up to you, what kind of outfit would you rather be in?

	Neurotic patients	*Cross section*
Combat overseas	35%	49%
Noncombat overseas	16	21
An outfit that will stay in U.S.	49	30
	100%	100%

13. WORRYING

Summary score for area

The worrying score, as its name implies, measures the extent to which a person worries about himself and the future, with a high score indicating a relatively unworried person.

Score	*Neurotic patients*	*Cross section*
4	25%	50%
3	26	25
2	24	15
1, 0	25	10
	100%	100%

Do you worry very much about things that might happen to you?

	Neurotic patients	*Cross section*
Not at all	16%	26%
Not very much	33	55
Quite a bit	27	14
A great deal	24	5
	100%	100%

When you get a very important job to do, do you worry about whether you will do all right or not?

	Neurotic patients	*Cross section*
Not at all	12%	18%
A little	38	56
A lot	50	26
	100%	100%

Do you ever worry about whether you will be injured in combat before the war is over?

	Neurotic patients	Cross section
Never	39%	41%
Hardly ever	29	36
Fairly often	22	17
A great deal	10	6
	100%	100%

Do you ever seriously worry about whether or not there will be a real depression after this war?

	Neurotic patients	Cross section
Almost never	37%	36%
Sometimes	41	43
Very often	22	21
	100%	100%

14. OVERSENSITIVITY

Summary score for area

The area of oversensitivity is concerned with such things as being easily offended, resentful of criticism, irritable, etc., a high score indicating an absence of the traits grouped together as oversensitivity.

Score	Neurotic patients	Cross section
10, 9, 8	29%	53%
7, 6	25	28
5, 4	23	13
3, 2, 1, 0	23	6
	100%	100%

Do you often say things you later wish you hadn't said?

	Neurotic patients	Cross section
Almost never	20%	28%
Sometimes	54	60
Very often	26	12
	100%	100%

How often do people get on your nerves so that you want to do just the opposite of what they want you to do?

	Neurotic patients	*Cross section*
Almost never	23%	37%
Sometimes	46	50
Very often	31	13
	100%	100%

Do you ever feel that people in the Army criticize you too much without any good reason?

	Neurotic patients	*Cross section*
Almost never	28%	42%
Only once in a while	34	41
Fairly often	21	11
Nearly all the time	17	6
	100%	100%

How often does it make you sore to have people tell you what to do?

	Neurotic patients	*Cross section*
Almost never	24%	29%
Sometimes	47	53
Very often	29	18
	100%	100%

Do you ever go out of your way to make things tough and unpleasant for somebody you don't like?

	Neurotic patients	*Cross section*
Almost never	74%	81%
Sometimes	22	17
Very often	4	2
	100%	100%

Do you ever feel like smashing things for no good reason?

	Neurotic patients	*Cross section*
Almost never	43%	71%
Sometimes	37	25
Very often	20	4
	100%	100%

How much does it bother you when you are ordered to do something you don't see a good reason for doing?

	Neurotic patients	*Cross section*
Not at all	10%	9%
Very little	17	24
Some	34	37
A great deal	39	30
	100%	100%

Did people seem to think you had a hot temper as a kid?

	Neurotic patients	*Cross section*
No	40%	58%
Undecided	16	10
Yes	44	32
	100%	100%

Do you find that you often have to tell people to mind their own business?

	Neurotic patients	*Cross section*
Almost never	32%	51%
Sometimes	48	42
Very often	20	7
	100%	100%

How often do people hurt your feelings?

	Neurotic patients	*Cross section*
Almost never	20%	44%
Sometimes	52	50
Very often	28	6
	100%	100%

15. PERSONAL ADJUSTMENT

Summary score for area

In this area are grouped questions dealing with self-confidence, and lack of self-pity, depression and anxiety. A high score represents possession of the traits just enumerated.

Score	*Neurotic patients*	*Cross section*
6, 5, 4	8%	33%
3, 2	25	37
1	23	18
0	44	12
	100%	100%

In general, how would you say you feel most of the time, in good spirits or in low spirits?

	Neurotic patients	*Cross section*
Good spirits	11%	37%
Good spirits sometimes; low spirits sometimes	44	50
Low spirits	45	13
	100%	100%

Are you ever worried and upset?

	Neurotic patients	*Cross section*
Hardly ever	11%	37%
Sometimes	34	48
Often	55	15
	100%	100%

Do you feel that you get more than your share of bad luck?

	Neurotic patients	*Cross section*
Almost never	19%	32%
Sometimes	51	52
Very often	30	16
	100%	100%

In general, what sort of time do you have in the Army?

	Neurotic patients	*Cross section*
Pretty good	14%	24%
It's about fifty-fifty	48	57
Pretty rotten time	38	19
	100%	100%

Do you ever get so blue and discouraged that you wonder whether anything is worth while?

	Neurotic patients	*Cross section*
Hardly ever	15%	36%
Not so often	23	35
Pretty often	31	17
Very often	31	12
	100%	100%

Do you think you can make good in the Army?

	Neurotic patients	Cross section
Very sure I can make good	14%	49%
I think so but I'm not sure	12	19
Undecided	31	20
Don't think I can make good	43	12
	100%	100%

16. PSYCHOSOMATIC COMPLAINTS

Summary scores in this area are graphed in Chart I in Part 1.

(1) Do you have any particular physical or health problem?[26]

	Neurotic patients	Cross section
2. No	8%	57%
3. Undecided	7	8
1. Yes	85	35
	100%	100%

(2) Do you often have trouble in getting to sleep or staying asleep?

	Neurotic patients	Cross section
3. Almost never	7%	48%
2. Sometimes	33	39
1. Often	60	13
	100%	100%

(3) Do your hands ever tremble enough to bother you?

	Neurotic patients	Cross section
3. Never	20%	68%
2. Sometimes	39	25
1. Often	41	7
	100%	100%

[26] The number in parenthesis, referring to the question number, and the numbers opposite the subcategories will facilitate subsequent cross reference from Tables 9 and 10 in Part 3.

(4) Have you ever had any fainting spells?

	Neurotic patients	*Cross section*
3. Never	38%	78%
2. A few times	41	18
1. Several times	21	4
	100%	100%

(5) Are you ever bothered by nervousness?

	Neurotic patients	*Cross section*
3. Never	5%	35%
2. Sometimes	25	48
1. Often	70	17
	100%	100%

(6) Have you ever been bothered by your heart beating hard?

	Neurotic patients	*Cross section*
3. Never	14%	46%
2. A few times	41	41
1. Often	45	13
	100%	100%

(7) Have you ever been bothered by pressures or pains in the head?

	Neurotic patients	*Cross section*
3. Never	16%	50%
2. Sometimes	37	37
1. Often	47	13
	100%	100%

(8) Have you ever had spells of dizziness?

	Neurotic patients	*Cross section*
3. Never	10%	44%
2. A few times	45	45
1. Many times	45	11
	100%	100%

(9) Do you ever bite your fingernails now?

	Neurotic patients	*Cross section*
3. Never	47%	68%
2. Sometimes	30	22
1. Often	23	10
	100%	100%

(10) Have you ever been bothered by shortness of breath when you were not exercising or working hard?

	Neurotic patients	*Cross section*
3. Never	19%	57%
2. Sometimes	41	31
1. Often	40	12
	100%	100%

(11) Are you ever troubled by your hands sweating so that they feel damp and clammy?

	Neurotic patients	*Cross section*
3. Never	10%	39%
2. Sometimes	33	43
1. Often	57	18
	100%	100%

(12) Are you ever troubled with sick headaches?

	Neurotic patients	*Cross section*
3. Never	14%	52%
2. Sometimes	44	38
1. Often	42	10
	100%	100%

(13) How often are you bothered by having an upset stomach?

	Neurotic patients	*Cross section*
4. Never	5%	22%
3. Not very often	37	59
2. Pretty often	38	15
1. Nearly all the time	20	4
	100%	100%

(14) Are you ever bothered by having nightmares (dreams that frighten or upset you very much)?

	Neurotic patients	*Cross section*
3. Never	20%	64%
2. A few times	49	32
1. Many times	31	4
	100%	100%

(15) Have you ever been troubled by "cold sweats"?

	Neurotic patients	*Cross section*
3. Never	20%	58%
2. A few times	54	37
1. Often	26	5
	100%	100%

PART 3

METHOD OF DERIVING WEIGHTS

The method used in deriving weights from the multiple regression of the criterion on the fifteen items in the initial test battery was an iterative technique.[27] From the point of view of sampling error, this method has an advantage over the Doolittle and related methods in that the iterations could be stopped at any point where it was apparent that chance was entering too much. Furthermore, it was far easier to compute. The basic data are presented in Table 9. In this table each entry represents a joint occurrence of two categories, that is, the proportion who answered the two indicated categories simultaneously.[28] Thus a diagonal entry repre-

[27] See Louis Guttman, "An Iterative Method for Multiple Correlation," in P. Horst, et al., *The Prediction of Personal Adjustment* (Social Science Research Council, New York, 1941), pp. 313–318.

[28] A word should be said about how the cross section and the psychoneurotic samples were weighted in constructing this table. As was indicated earlier, the proper weighting to maximize the discrimination was one which would equalize the variances of the two groups' prediction scores. Since on other scorings of the psychosomatic complaints, the scores of the psychoneurotic group were noted to have smaller total variances than those of the cross section, the former group was given the heavier weight. It was estimated, without bothering to solve the equations for the proper weights, that a ratio of 9 to 1 would be about right, and this is the weighting used in Table 9. As it turned out, this was something of an overweighting of the psychoneurotic group, but it made little difference. The proper ratio should be less than 9 to 1 and more than 1 to 1, but would, in any case, not have given different results from those obtained, as far as total discrimination is concerned.

sents the proportion answering the indicated category (since it always occurs with itself); two categories within the same question have a zero joint occurrence, since they are mutually exclusive; and two categories from different questions have relative frequencies bigger than zero. The sum of each three-by-three square in the table, where the rows belong all to one question and the columns belong all to another question, is unity.

For the first iteration, zero-order weights were used, that is, the least square weights that an item would be given if it were to be the only predictor. These are quite simple to compute for qualitative items, being simply the ratio of the joint occurrence of a category with the criterion to the total frequency of that category. Adding any constant to the weights of all the categories of a question simply changes each person's total score by exactly that constant, leaving the correlation with the criterion unchanged. Therefore, to simplify further computations, a constant equal to the weight of the top category was subtracted from the category weights of each question, thus automatically reducing the weight of the top category to zero. The work from then on proceeded with Table 10, which is basically Table 9 with the categories with zero weight eliminated. In Table 10, a dummy variable was added to represent the additive constant for the regression equation; that is, everybody had a score of unity on the dummy variable, so its diagonal entry is unity, and its nondiagonal entries are simply the diagonal entries of the corresponding categories.

If x_{ij} represents the entry in row i and column j of Table 9 and if c_i represents the entry in the ith row of column C, then the multiple regression weights, w_j, are defined to be those weights which satisfy the normal equations:

$$c_i - \sum_j w_j x_{ij} = 0$$

The iterative process first tests the trial weights w_1, to see to what extent they satisfy the normal equations. These weights are of course not in scale and must be divided by a constant. The column $\Sigma w_1 x$ in Table 10 when multiplied by a constant[29] to yield the column $k\Sigma w_1 x$ shows to what extent the zero-order weights are proportional to the multiple regression weights. It may be noted how close to zero are the entries in the last column of Table 10. If these were

[29] The constant employed is the slope of the line of best fit of the $\Sigma w_1 x$'s to the c's. In this case, $k = .231$.

TABLE 9

Joint Occurrences of Psychosomatic Complaints: Cross Section and Psychoneurotics Weighted One to Nine and Combined (January-February 1944)

Question*		1			2			3			4		
	Subcategory	*2*	*3*	*1*	*3*	*2*	*1*	*3*	*2*	*1*	*3*	*2*	*1*
1	2	.127			039	.050	038	.071	035	.021	086	.032	.010
	3		073		009	038	.026	.018	039	.015	.025	031	016
	1			.800	061	244	.495	157	301	342	.308	322	170
2	3	039	.009	.061	.109			078	022	.008	072	024	.013
	2	.050	.038	.244		.332		.100	149	083	.152	.134	047
	1	.038	.026	.495			.559	069	203	288	.196	227	.136
3	3	.071	.018	.157	078	.100	.069	247			153	071	023
	2	.035	.039	301	.022	.149	203		.374		.135	166	074
	1	.021	.015	.342	008	.083	.288			.379	.130	.149	099
4	3	.086	025	.308	.072	.152	.196	.153	.135	.130	418		
	2	032	.031	322	024	134	227	071	166	.149		386	
	1	010	.016	170	.013	047	136	023	074	.099			.196
5	3	038	.004	.036	039	032	.007	.066	.011	002	061	012	005
	2	047	039	187	041	134	099	119	136	019	141	099	.034
	1	042	.029	.578	.029	.166	452	063	228	358	217	.275	157
6	3	048	016	105	046	.061	062	092	048	029	102	052	015
	2	051	040	319	042	.156	.211	.100	167	142	174	178	.059
	1	029	017	.376	021	.115	287	.056	159	207	143	.156	123
7	3	056	011	127	.053	.084	.057	099	048	.047	.121	.059	.014
	2	.048	.024	.294	036	143	187	.087	.160	.119	.157	.148	.060
	1	024	.038	379	020	.106	.315	061	.166	.213	.140	179	.122
8	3	041	.009	.082	.045	.063	.026	.074	033	.026	110	017	.006
	2	.060	038	.353	049	.175	225	116	185	148	.212	.192	.047
	1	.026	.026	.365	.015	.095	307	.057	156	204	.098	.177	143
9	3	078	.029	379	074	.185	227	172	.177	.137	.249	168	069
	2	038	022	233	.023	.098	173	052	.116	125	102	.123	069
	1	.011	.021	.188	.012	.049	159	.023	.082	.116	068	.094	.059
10	3	.062	019	.148	.065	.080	.084	116	.073	039	.140	064	.026
	2	040	031	329	028	144	.228	.073	.175	152	.175	.157	.066
	1	.025	.022	.322	016	.108	.247	058	127	.186	.102	164	.104
11	3	.046	.015	.068	037	055	.037	.075	034	.020	.072	.035	.022
	2	048	025	263	.049	.127	.161	097	.154	085	157	.125	055
	1	.033	.032	.470	.023	.150	.362	.074	.188	.273	.190	225	.120
12	3	.054	.010	.115	.057	078	044	101	045	.032	.123	041	.014
	2	056	.038	.342	037	.164	234	.094	.184	.157	.188	.174	072
	1	017	.026	.344	.015	.090	.281	052	.146	.189	.107	170	.110
13	4	.021	005	.044	.024	.024	.023	.036	.022	.012	.049	014	.007
	3	073	032	.285	055	.176	.160	.141	.143	.108	.187	144	.060
	2,1	.033	.036	.470	030	.133	376	.071	.210	257	.183	.227	.128
14	3	.060	.019	.163	070	.098	.075	.131	.077	036	.158	.062	.022
	2	.048	.031	398	.032	.184	260	.103	.198	.175	.193	.209	.075
	1	.019	.022	.240	.007	.050	.224	.013	.100	.167	067	.115	.099
15	3	.065	.019	.153	.061	.099	078	.115	.078	.044	.153	.062	.022
	2	.052	.045	.422	.041	.183	295	.107	.225	.187	.211	.210	.098
	1	.010	.009	.224	.007	.050	.187	.025	.072	.147	.054	.113	.077

* The exact wordings of questions and subcategories are shown in Part 2 of this chapter. In that part, the questions here referred to by number are indicated by a parenthetical number preceding the question. The subcategories in this table follow the numbering used in Part 2.

TABLE 9 (Continued)

Question		*5*			*6*			*7*			*8*		
	Subcategory	*3*	*2*	*1*	*3*	*2*	*1*	*3*	*2*	*1*	*3*	*2*	*1*
	2	.038	.047	.042	048	.051	.029	.056	.048	.024	.041	060	.026
1	3	004	.039	.029	.016	040	.017	.011	.024	.038	.009	.038	.026
	1	.036	187	578	105	.319	.376	.127	.294	379	.082	353	.365
	3	039	.041	.029	.046	.042	.021	.053	.036	.020	045	.049	015
2	2	.032	.134	.166	.061	.156	.115	084	143	.106	.063	.175	.095
	1	007	.099	.452	.062	.211	.287	057	.187	.315	.026	.225	307
	3	066	119	.063	.092	.100	.056	.099	087	.061	.074	.116	.057
3	2	011	136	.228	.048	.167	159	.048	.160	.166	.033	.185	.156
	1	.002	.019	.358	.029	.142	207	.047	.119	.213	.026	.148	.204
	3	.061	.141	.217	102	.174	.143	.121	.157	.140	.110	.212	.098
4	2	.012	.099	.275	.052	.178	.156	.059	.148	.179	.017	.192	.177
	1	.005	.034	.157	015	.059	.123	.014	.060	.122	.006	.047	.143
	3	078			045	.023	.010	.049	.016	.013	.041	026	011
5	2		273		072	.136	066	.063	133	.078	052	.151	.071
	1			649	.052	.251	346	082	.217	.349	.041	.273	.334
	3	.045	072	052	.169			074	.054	040	051	.085	.032
6	2	.023	.136	.251		.409		087	.184	138	051	.233	.126
	1	.010	066	.346			.422	.033	.126	262	.032	.132	259
	3	049	.063	.082	074	.087	.033	194			.083	.086	.025
7	2	016	.133	217	.054	184	.126		.366		038	.210	.118
	1	013	.078	.349	040	.138	262			440	.012	154	275
	3	041	.052	041	.051	.051	032	083	.038	.012	133		
8	2	026	151	.273	.085	.233	.132	.086	.210	.154		.450	
	1	.011	.071	.334	032	126	.259	025	.118	.275			.417
	3	.065	.159	261	.104	.207	.175	127	.180	.180	.085	.225	.177
9	2	009	081	204	048	111	135	.041	.110	.143	.036	.129	.129
	1	003	034	.183	017	092	.110	.027	076	.117	.012	097	112
	3	051	088	090	087	084	058	.086	.086	.057	.067	.109	.052
10	2	021	119	260	057	.199	.145	.062	.163	.175	.047	.202	.151
	1	.006	067	.298	025	.126	219	047	.117	.207	.019	.138	.213
	3	035	.058	036	058	044	027	.058	.037	033	.038	.061	030
11	2	030	126	180	.058	173	105	072	148	.117	057	163	117
	1	013	090	432	054	191	289	064	181	290	039	226	270
	3	048	073	057	063	.074	041	105	053	020	077	074	027
12	2	022	143	.270	070	.186	179	075	.242	.119	.047	240	.147
	1	008	.057	.322	.037	.149	201	015	.070	302	009	.135	242
	4	.020	026	.025	029	028	013	038	018	.014	.027	.032	.011
13	3	.046	.148	197	.090	.164	.137	.095	138	.158	.073	.195	.124
	2,1	012	101	426	050	.218	271	.061	.209	.269	033	.223	282
	3	.054	.111	077	.081	.101	061	087	.087	.068	.077	.119	.046
14	2	.020	.132	.324	.075	.216	.185	.090	.180	.206	.046	.236	.194
	1	.003	031	.247	.014	.092	.174	.017	.099	.166	.010	.094	177
	3	.045	.098	.094	.077	.099	.061	.088	086	062	.072	.115	.050
15	2	028	.154	.337	.082	.239	.198	.080	.216	.224	.051	258	.209
	1	.005	.022	.217	.010	.072	.163	.026	.064	.154	.010	076	.157

TABLE 9 (Continued)

Question		9			10			11			12		
	Subcategory	3	2	1	3	2	1	3	2	1	3	2	1
1	2	078	.038	011	062	.040	025	.046	.048	033	.054	.056	.017
	3	029	022	.021	019	031	022	015	.025	.032	010	038	.026
	1	379	233	188	148	329	.322	.068	.263	.470	115	342	.344
2	3	074	023	012	065	.028	.016	.037	.049	.023	057	.037	015
	2	.185	098	049	.080	.144	.108	055	.127	.150	.078	.164	090
	1	227	.173	.159	084	.228	.247	037	.161	.362	044	.234	.281
3	3	.172	052	023	116	.073	.058	075	.097	074	.101	.094	052
	2	177	116	082	.073	.175	.127	034	154	.188	.045	.184	.146
	1	137	.125	116	039	.152	.186	.020	085	.273	032	.157	.189
4	3	249	102	.068	.140	.175	.102	.072	.157	.190	.123	188	.107
	2	.168	123	094	.064	.157	.164	035	.125	225	.041	174	170
	1	069	069	059	026	066	.104	.022	055	120	014	072	110
5	3	065	.009	003	051	.021	.006	.035	.030	013	048	.022	008
	2	159	.081	034	.088	.119	.067	.058	.126	090	.073	143	057
	1	.261	.204	.183	090	.260	.298	.036	.180	432	.057	270	.322
6	3	.104	048	017	087	057	025	.058	058	054	.063	070	.037
	2	207	111	092	084	.199	.126	.044	.173	.191	.074	186	149
	1	.175	.135	110	.058	.145	.219	.027	.105	289	041	.179	201
7	3	.127	041	027	086	.062	.047	.058	072	064	.105	075	015
	2	180	.110	076	086	.163	.117	.037	.148	.181	053	242	070
	1	.180	143	.117	057	.175	.207	.033	.117	290	020	.119	.302
8	3	085	036	.012	.067	.047	.019	.038	.057	.039	.077	047	009
	2	.225	.129	097	109	202	.138	061	.163	.226	074	240	.135
	1	177	.129	.112	.052	.151	213	.030	.117	.270	027	147	242
9	3	.486			.150	.193	143	087	179	.221	116	221	150
	2		.294		049	.129	116	.026	.104	.164	040	131	123
	1			.220	030	078	112	017	053	.150	022	083	114
10	3	.150	.049	.030	229			066	085	078	091	083	054
	2	.193	.129	.078		.400		.039	156	.205	.052	201	148
	1	.143	.116	.112			.371	024	095	252	.035	152	185
11	3	.087	026	017	066	039	024	.129			056	.056	017
	2	.179	104	053	085	.156	.095		336		068	155	113
	1	221	.164	150	.078	.205	.252			.535	055	224	.257
12	3	116	040	022	.091	.052	.035	056	.068	.055	178		
	2	221	.131	083	.083	201	152	056	155	224		.435	
	1	.150	.123	.114	054	.148	.185	.017	113	257			.387
13	4	.047	.010	.013	.036	.022	.013	031	027	.012	038	021	011
	3	206	.118	068	111	.170	110	.067	158	.167	.096	190	104
	2,1	232	.165	140	.082	209	248	032	150	356	044	.223	271
14	3	159	053	031	.111	.085	.047	072	092	079	091	.101	051
	2	.243	.138	095	.091	.225	161	045	.185	.246	072	.228	.177
	1	.084	.103	094	027	092	.162	.011	060	210	015	.106	.160
15	3	.152	053	.032	103	.081	053	079	087	071	092	.095	050
	2	250	.165	.104	099	246	174	039	.214	.266	.063	.255	.201
	1	.084	.076	083	027	.073	.144	011	.035	.198	022	.086	136

TABLE 9 (Continued)

Question	*Subcategory*	*13* 4	*13* 3	*13* 2,1	*14* 3	*14* 2	*14* 1	*15* 3	*15* 2	*15* 1	*Joint occurrence with criterion (C)*	*Zero-order weights* *Ratio of occurrence with criterion to total frequency*	*Zero-order weights* *Corrected to have zero origin (w_1)*
1	2	.021	.073	033	060	048	019	.065	.052	.010	.070	.553	.000
	3	.005	.032	036	.019	031	022	.019	.045	.009	.067	.911	.358
	1	.044	285	.470	.163	.398	.240	.153	.422	.224	.763	.954	.401
2	3	024	.055	.030	.070	032	.007	061	.041	007	061	.555	.000
	2	.024	.176	133	.098	.184	050	.099	183	.050	.293	.884	.329
	1	023	.160	.376	.075	.260	224	.078	.295	.187	.546	976	.421
3	3	.036	141	071	.131	.103	013	115	.107	.025	.179	.724	.000
	2	.022	143	210	.077	.198	100	.078	.225	072	.349	.933	.209
	1	.012	.108	.257	.036	.175	.167	044	.187	.147	.372	.982	.258
4	3	.049	187	.183	.158	193	.067	.153	211	054	.340	814	.000
	2	014	.144	227	.062	209	.115	062	210	113	.367	.953	.139
	1	007	060	.128	022	075	099	022	098	077	193	.980	.166
5	3	020	046	012	.054	020	003	045	028	005	042	544	000
	2	026	.148	101	.111	132	031	.098	.154	022	.226	.826	.282
	1	.025	.197	.426	077	.324	247	.094	.337	.217	.632	.974	.430
6	3	029	.090	.050	.081	075	014	.077	082	010	.123	.731	.000
	2	.028	.164	.218	.101	216	.092	.099	239	072	.368	.899	.168
	1	.013	.137	.271	061	185	.174	061	.198	163	.409	968	.237
7	3	.038	095	.061	.087	.090	.017	.088	.080	.026	.144	741	.000
	2	.018	.138	.209	.087	.180	099	.086	216	.064	329	900	.159
	1	.014	.158	.269	.068	.206	166	.062	224	.154	.427	.970	.229
8	3	.027	.073	.033	077	.046	.010	.072	.051	010	.090	.674	.000
	2	.032	.195	.223	.119	.236	.094	.115	258	.076	.405	900	.226
	1	.011	.124	282	.046	194	.177	.050	.209	.157	.405	972	.298
9	3	.047	.206	232	159	.243	.084	.152	250	084	418	861	.000
	2	010	.118	165	.053	.138	.103	.053	.165	076	.272	.925	.064
	1	.013	068	.140	.031	.095	094	032	.104	083	.210	953	.092
10	3	.036	.111	082	.111	.091	.027	.103	.099	027	.172	.750	.000
	2	.022	.170	.209	.085	.225	092	081	.246	073	.370	.924	.174
	1	.013	.110	248	.047	.161	.162	.053	174	.144	358	.967	.217
11	3	031	.067	.032	.072	.045	011	.079	039	.011	090	.697	.000
	2	.027	.158	.150	.092	.185	.060	.087	214	035	.293	.874	.177
	1	.012	.167	.356	.079	246	210	071	.266	.198	517	.965	.268
12	3	038	.096	.044	.091	.072	.015	.092	063	.022	.126	.709	.000
	2	.021	.190	.223	.101	228	.106	.095	.255	.086	.398	.914	.205
	1	.011	.104	.271	.051	.177	.160	.050	.201	.136	.376	.972	.263
13	4	.070			048	.019	.003	.041	.025	.005	.049	.691	.000
	3		.392		.119	.210	.062	.123	.221	.048	.332	.848	.157
	2,1			.538	.076	.247	.216	.074	.273	.191	.519	.965	.274
14	3	.048	.119	.076	.243			106	.105	.031	.179	.738	.000
	2	.019	.210	.247		.476		112	288	.076	.445	.934	.196
	1	.003	.062	216			.281	.019	.125	.137	.276	.983	.245
15	3	.041	.123	.074	.106	.112	.019	237			.179	.755	.000
	2	.025	.221	.273	.105	.288	.125		.519		.482	.928	.173
	1	.005	.048	191	031	.076	.137			.244	.239	.981	.226

TABLE 10

DERIVATION OF REGRESSION WEIGHTS FROM JOINT OCCURRENCES OF PSYCHOSOMATIC COMPLAINTS: CROSS SECTION AND PSYCHONEUROTICS WEIGHTED ONE TO NINE AND COMBINED (January-February 1944)

Question		Dummy variable	1		2		3		4		5		6		7		8		9	
	Sub-category		3	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1
Dummy variable		1 000	073	.800	.332	.559	.374	.379	386	.196	273	.649	.409	.422	366	.440	.450	417	294	.220
1	3	.073	.073		038	.026	039	015	.031	.016	.039	029	040	.017	024	.038	038	.026	.022	.021
	1	.800		.800	244	.495	.301	.342	322	.170	187	578	.319	.376	.294	379	353	.365	.233	.188
2	2	.332	.038	.244	.332		.149	.083	.134	.047	134	.166	.156	.115	.143	.106	.175	.095	.098	.049
	1	.559	.026	.495		.559	.203	.288	227	.136	.099	.452	.211	.287	187	315	.225	.307	.173	.159
3	2	.374	.039	.301	.149	.203	.374		.166	.074	.136	.228	167	.159	160	166	.185	.156	.116	082
	1	.379	.015	.342	.083	288		.379	149	.099	.019	358	.142	.207	119	.213	.148	.204	.125	.116
4	2	.386	.031	.322	.134	227	.166	.149	386		.099	.275	178	.156	.148	179	192	.177	.123	094
	1	.196	.016	.170	.047	.136	.074	.099		.196	034	157	.059	.123	060	.122	047	.143	.069	059
5	2	.273	.039	.187	.134	.099	136	.019	099	034	273		.136	.066	133	078	151	071	081	.034
	1	.649	.029	.578	.166	.452	.228	.358	275	.157		.649	.251	346	217	349	.273	334	.204	.183
6	2	.409	040	319	.156	.211	.167	142	.178	.059	136	251	409		184	.138	.233	126	.111	092
	1	.422	.017	.376	115	287	.159	.207	.156	.123	066	346		422	.126	.262	.132	259	.135	.110
7	2	.366	.024	.294	.143	.187	.160	119	.148	.060	.133	217	.184	126	366		.210	.118	110	.076
	1	.440	.038	.379	106	.315	.166	213	179	122	078	349	.138	.262		.440	.154	.275	.143	.117
8	2	450	038	.353	.175	.225	.185	148	.192	047	.151	273	233	.132	210	.154	450		.129	.097
	1	.417	.026	.365	.095	307	.156	204	.177	.143	.071	334	.126	259	118	275		.417	.129	.112
9	2	.294	.022	.233	098	.173	.116	.125	123	069	.081	204	111	135	110	.143	.129	.129	.294	
	1	.220	.021	.188	049	159	.082	116	094	059	034	.183	092	110	076	117	097	.112		220
10	2	400	.031	.329	144	.228	.175	152	157	066	.119	260	.199	.145	163	175	.202	151	.129	078
	1	.371	022	322	.108	247	.127	186	164	.104	.067	298	126	219	117	207	138	213	.116	112
11	2	.336	025	263	127	.161	154	085	.125	.055	126	.180	.173	.105	148	.117	163	.117	.104	053
	1	.535	032	.470	.150	362	188	273	225	120	.090	.432	.191	289	181	290	226	270	.164	150
12	2	435	038	.342	164	234	.184	.157	174	.072	143	.270	.186	.179	242	.119	.240	.147	131	.083
	1	.387	026	344	.090	.281	146	.189	170	110	057	322	149	.201	.070	302	.135	.242	.123	.114
13	3	.392	.032	285	.176	.160	143	108	144	060	.148	.197	.164	137	.138	158	195	.124	.118	068
	2,1	.538	036	.470	133	376	210	257	227	128	.101	426	218	271	209	269	.223	.282	.165	.140
14	2	.476	031	.398	.184	260	.198	175	209	075	132	324	216	.185	180	206	.236	194	138	095
	1	.281	022	.240	.050	.224	.100	.167	115	.099	031	247	092	174	099	.166	.094	.177	.103	.094
15	2	519	.045	.422	183	.295	.225	187	.210	098	154	337	239	198	216	224	258	209	.165	.104
	1	.244	009	224	050	.187	.072	147	.113	077	022	.217	072	163	.064	154	076	.157	076	.083

TABLE 10 (Continued)

Question	Subcategory	Dummy variable	10: 2	10: 1	11: 2	11: 1	12: 2	12: 1	13: 3	13: 2,1	14: 2	14: 1	15: 2	15: 1	C	w_{1i}	$\Sigma w_{1j}x$	$k\Sigma w_{1j}x$	$c_i - k\Sigma w_{1j}x$
Dummy variable		1 000	.400	371	.336	535	435	387	.392	538	476	.281	.519	.244	.900	900	3.862	.892	.008
1	3	.073	031	.022	025	032	038	026	032	036	031	.022	045	.009	067	.358	0 275	.064	.003
	1	800	329	.322	.263	470	.342	344	.285	.470	398	240	422	.224	.763	.401	3 246	750	.013
2	2	.332	.144	108	.127	150	164	090	176	.133	.184	050	.183	.050	.293	329	1 210	.280	.013
	1	559	228	.247	161	362	.234	281	160	376	.260	224	295	187	.546	421	2 380	.550	−.004
3	2	.374	.175	.127	154	188	184	.146	.143	210	.198	.100	.225	072	349	209	1 499	346	.003
	1	379	.152	186	.085	.273	157	189	108	257	175	167	187	.147	.372	258	1 643	.379	− 007
4	2	386	.157	164	125	225	174	170	.144	227	209	115	210	.113	.367	.139	1 585	.366	.001
	1	.196	066	.104	.055	120	072	.110	.060	.128	.075	.099	098	077	.193	166	0 852	.197	−.004
5	2	.273	119	067	.126	090	143	.057	148	101	.132	031	.154	.022	.226	.282	0 934	.216	.010
	1	.649	.260	.298	.180	432	.270	322	197	.426	.324	.247	.337	.217	.632	.430	2.760	.638	−.006
6	2	.409	199	.126	173	191	186	149	164	218	216	.092	239	.072	.368	.168	1.567	.362	.006
	1	.422	.145	219	.105	.289	.179	.201	.137	271	.185	174	.198	.163	.409	.237	1.802	.416	−.007
7	2	.366	.163	.117	.148	.181	.242	.070	.138	.209	.180	.099	.216	.064	.329	.159	1 409	.325	.004
	1	.440	.175	.207	117	290	.119	.302	.158	.269	.206	.166	.224	.154	427	.229	1.888	.436	−.009
8	2	.450	.202	.138	163	.226	.240	.135	.195	.223	.236	.094	.258	.076	.405	.226	1.703	.393	.012
	1	.417	.151	.213	.117	.270	.147	.242	.124	.282	.194	.177	.209	.157	.405	.298	1.805	.417	−.012
9	2	.294	.129	.116	.104	.164	131	.123	.118	.165	.138	.103	.165	.076	272	.064	1.184	.273	−.001
	1	.220	.078	.112	.053	.150	.083	.114	.068	.140	.095	094	.104	.083	.210	.092	0 941	.217	−.007
10	2	.400	.400		.157	205	.201	.148	.170	.209	.225	.092	246	.073	.370	.174	1.583	.366	.004
	1	.371		.371	.095	.252	.152	185	.110	.248	161	.162	.174	.144	.358	.217	1.587	.367	−.009
11	2	.336	.156	.095	.336		.155	.113	.158	.150	.185	.060	.214	.035	.293	.177	1.238	.286	.007
	1	.535	.205	.252		.535	.224	.257	.167	.356	.246	210	266	.198	.517	268	2.269	.524	−.007
12	2	.435	.201	.152	.155	.224	.435		.190	.223	.228	.106	.255	.086	.398	.205	1.690	.390	.008
	1	.387	.148	.185	.113	.257		.387	.104	.271	.177	.160	.201	.136	.376	.263	1.680	.388	−.012
13	3	.392	.170	.110	.158	.167	.190	.104	392		.210	.062	.221	048	.332	.157	1.390	.321	.011
	2,1	.538	.209	.248	.150	.356	.223	.271		.538	.247	.216	.273	.191	.519	.274	2.285	528	−.009
14	2	.476	.225	.161	.185	.246	.228	177	210	.247	.476		.288	.076	.445	.196	1.886	.436	.009
	1	.281	.092	.162	.060	.210	.106	.160	.062	.216		.281	.125	.137	.276	.245	1.250	289	−.013
15	2	.519	.246	.174	.214	.266	.255	201	.221	.273	.288	125	.519		.482	.173	2.060	.476	.006
	1	.244	.073	.144	.035	.198	.086	.136	.048	.191	.076	.137		.244	.239	.226	1.082	.250	−.011

exactly zero, then the w_1 would be precisely proportional to the multiple regression weights. As it is, they are so close to being proportional that no further iterations were tried. Little gain in discrimination could have resulted from continuing the iterative

TABLE 11

A TEST OF THE PROPORTIONALITY OF THE SIMPLE TRICHOTOMOUS WEIGHTING OF PSYCHOSOMATIC COMPLAINTS ITEMS TO THEIR MULTIPLE REGRESSION WEIGHTS

Question	*Subcategory*	C	$\Sigma w_2 x$	$k\Sigma w_2 x$*	$c - k\,\Sigma w_2 x$
Dummy		.900	19.389	.873	.027
1	3	.067	1.293	.058	.009
	1	.763	16.619	.748	.015
2	2	.293	5.641	.254	.039
	1	.546	12.592	.567	−.021
3	2	.349	7.366	.331	.018
	1	.372	8.884	.400	−.028
4	2	.367	8.047	.362	.005
	1	.193	4.720	.212	−.019
5	2	.226	4.194	.189	.037
	1	.632	14.554	.655	−.023
6	2	.368	7.571	.341	.027
	1	.409	9.697	.436	−.027
7	2	.329	6.803	.306	.023
	1	.427	10.153	.457	−.030
8	2	.405	8.149	.367	.038
	1	.405	9.769	.440	−.035
9	2	.272	6.074	.273	−.001
	1	.210	5.166	.232	−.022
10	2	.370	7.743	.348	.022
	1	.358	8.586	.386	−.028
11	2	.293	5.861	.264	.029
	1	.517	12.022	.541	−.024
12	2	.398	8.191	.369	.029
	1	.376	9.100	.410	−.034
13	3	.332	6.564	.295	.037
	2,1	.519	12.042	.542	−.023
14	2	.445	9.255	.416	.029
	1	.276	6.936	.312	−.036
15	2	.482	10.181	.458	.024
	1	.239	6.036	.272	−.033

* $k = .045$

process, while sampling error might have been increased. But the only conclusion possible from the fact that the iterative process could not be carried beyond the first iteration was that differential weightings of the items would not yield any substantial improvement in discrimination over that obtained by simple weightings. As a further test of this conclusion, the simple trichotomous weights were used for the first iteration of the data in Table 10, and they also proved to be almost proportional to the multiple regression weights. The data basic to this conclusion are shown in Table 11. (The weights here are in reverse order to those used elsewhere in Chapters 13 and 14, with a weight of 2 given the subcategory which occurred most disproportionately among the psychoneurotics.)

In summary, then, almost identical results were obtained when three different methods of weighting the items in the psychosomatic complaints score were tried:

	PERCENTAGES RECEIVING CRITICAL SCORES	
Method of weighting used	*Cross section*	*Psychoneurotic sample*
Simple dichotomous (0, 1)	28.6%	89.4%
Simple trichotomous (0, 1, 2)	26.8	89.6
Differential trichotomous	28.6	90.8

Considerations deriving from the fact that these items formed a quasi scale led us to believe that the simple scorings approximated the results to be obtained by employing multiple regression techniques, and an empirical investigation of the interrelationships confirmed this assumption. In view of the almost identical results obtained, the simple trichotomous weights were selected for practical use, since they were the simplest to score.

CHAPTER 14

THE SCREENING OF PSYCHONEUROTICS: COMPARISON OF PSYCHIATRIC DIAGNOSES AND TEST SCORES AT ALL INDUCTION STATIONS[1]

IN JULY 1945, all induction stations in the United States were instructed to keep a daily record of the NSA scores of all men receiving preinduction physical examinations during the month of August 1945, the men being classified into one of seven groups according to disposition.[2] These categories were a major division of men into those found acceptable and those rejected for military service, the rejections being divided by cause into administrative, medical and physical, and psychiatric rejections. Within the latter group, a four-way division into major diagnostic categories also was secured, namely, psychoneurotic, psychopathic, psychotic, and all others. Thus records on the disposition of over one hundred thousand cases were obtained.

It must be kept in mind that the NSA was not intended to be a substitute for psychiatric examination. It was rather to be an adjunct or aid to psychiatric examination. Men were to take the test and the examining physician would have the test results in front of him to aid him in his diagnosis. If a man made a favorable score on the test, with no "critical signs," as described in the previous chapter, the examination, it was hoped, could be even more perfunctory than normally, thus giving the already overburdened physician more time for a somewhat more thorough than average examination of the doubtful cases screened by the test. In situations of unusual time pressure because of the volume of inductees, he might, at his discretion, dispense with examination in those cases who made favorable numerical scores on the test and showed

[1] By Shirley A. Star.

[2] Instructions were given in ASF Circular No. 281 (24 July 1945).

no "critical signs." Actually, during August 1945, the month in which the war ended, the rush of inductees was not great and, as far as is known, physicians nowhere exercised this option. How much they actually paid attention to the NSA scores we do not know. In spite of the fact that the test went out under the auspices of the Neuropsychiatric Division of the Office of the Surgeon General and was administered by trained psychologists, it would not be surprising if by a good many clinicians it was either treated with tolerant amusement or completely ignored, even though their own clinical examination hardly could be more than perfunctory in the majority of cases.

The fact, however, that the test scores were available to examining physicians, who were free to use them or discard them, means that there is a lack of experimental independence in our comparison of NSA scores and psychiatrists' judgments. In developing the test, as reported in the preceding chapter, experimental independence was obtained. When the test was tried out at induction stations, with the quite satisfactory results reported in the previous chapter, predictions were made about the psychiatrists' disposition of the cases, in a situation where psychiatrists had *no* knowledge of the test scores. In the nationwide study now being reviewed, we are in no position to say how much a knowledge of the test scores influenced the correlation between test scores and psychiatrists' ratings. But we do know that at all stations, irrespective of test scores, all men were given a personal psychiatric examination in addition to taking the test and the relationship between the test score and the clinical disposition is a matter which should be of very considerable scientific interest, even though the one may have influenced the other.

In order to evaluate the results, we shall look first at the facts as to diagnosis and disposition by examining physicians. Next we shall relate these diagnoses to the NSA scores made by the men so diagnosed. Finally, we shall make some special observations on the value of "critical signs" to supplement the NSA score.

Diagnosis and Disposition of Cases by Examining Physicians, August 1945

Table 1 shows that throughout the United States, in August 1945, 57.7 per cent of all literate preinduction examinees were found acceptable for service on medical grounds. The psychiatric rejects constituted 14.0 per cent, of whom the largest proportion were

diagnosed as psychoneurotic and the next largest proportion as psychopathic.[3]

However, as Table 2 makes abundantly clear, there was extremely wide variation in the psychiatric rejection rates. The range, among the fifty-five induction stations, was from 0.5 per cent at Camp Beale, California, to 50.6 per cent at Manchester, New Hampshire. Some of the percentages of rejections shown in Table 2 are based on a rather small number of cases, but even if we compare pairs of relatively large stations in the same geographical region we see that the variability is great.

TABLE 1

CLASSIFICATION OF LITERATE PREINDUCTION EXAMINEES IN AUGUST 1945

Classification	*Per cent in each category*		
Acceptable for service			57.7%
Rejected			
Administrative rejections		*1.2%*	
Medical and physical rejections (excluding psychiatric)		*27.1*	
Psychiatric rejections (including rejections for both psychiatric and other reasons)			
Psychoneurotic	*5.6%*		
Psychopathic	*3.9*		
Psychotic	*0.3*		
Other psychiatric	*4.2*	*14.0*	42.3
			100.0%

Not only was there wide variation in the psychiatric rejection rates, but also there was wide variation in the specific diagnoses given for these psychiatric rejects. This is shown in Table 3. While, in the nation as a whole, 39.9 per cent of all psychiatric rejects were diagnosed as psychoneurotic, the percentages varied, among stations with at least 50 rejects, all the way from 2.7 to 90.2. There were 29 stations in which the dominant tendency was to classify men as psychoneurotic, 16 in which the psychiatric rejections were most often psychopathic, and 10 in which they were usually called by neither of these labels.

It might be argued, by way of explaining such enormous variability in diagnosis, that the statistics in Tables 2 and 3 represent a faithful picture of the actual incidence among the populations

[3] Only data on literate inductees are considered in this chapter, since men who failed to pass a standard literacy test administered by the psychologists were not given the NSA, though they were of course examined by physicians.

TABLE 2

DISPOSITION OF PREINDUCTION EXAMINEES, CLASSIFIED BY SERVICE COMMAND AND INDUCTION STATION

(August 1945)

SERVICE COMMAND AND INDUCTION STATION	DISPOSITION IN PERCENTAGES: Accepted	Rejected: Administrative	Rejected: Medical and physical	Rejected: Psychiatric	TOTAL: Per cent	TOTAL: Number
Total	57.7	1.2	27.1	14.0	100.0	*107,385*
First Service Command	57.7	1.2	19.8	21.3	100.0	*5,707*
Boston (Fort Banks), Mass	61.0	0.9	19.5	18.6	100.0	*2,135*
Portland (Fort Preble), Me.	55.3	3.3	23.6	17.8	100.0	*517*
Manchester (Grenier Field), N.H.	36.3	2.0	11.1	50.6	100.0	*666*
New Haven, Conn.	60.2	0.1	22.3	17.4	100.0	*1,590*
Rutland, Vt.	63 5	2.4	15.7	18.4	100.0	*370*
Springfield, Mass.	64.1	0.5	24.0	11.4	100.0	*429*
Second Service Command	51.3	0.3	36.4	12.0	100.0	*15,252*
Albany-Syracuse, N.Y.	61.5	0 1	28.9	9.5	100.0	*1,411*
Buffalo-Rochester, N.Y.	60.6	0.3	29.0	10.1	100.0	*1,668*
Newark-Camden, N.J.	69.1	—	21.6	9.3	100.0	*2,690*
New York, N.Y.	43.2	0.3	43.0	13.5	100.0	*9,483*
Third Service Command	53.3	0.3	30.1	16.3	100.0	*16,828*
Baltimore, Md.	55.6	—	24.0	20.4	100.0	*2,987*
Harrisburg, Pa.	50.0	0.6	30.2	19.2	100.0	*1,789*
Philadelphia, Pa.	51.5	0.3	40.8	7.4	100.0	*3,483*
Pittsburgh, Pa	47.3	0.2	28.3	24.2	100.0	*4,587*
Richmond, Va.	71.3	0.7	23.1	4.9	100.0	*1,373*
Roanoke, Va.	54.9	0.1	23.7	21.3	100.0	*920*
Wilkes-Barre, Pa.	57.1	—	33.7	9.2	100.0	*1,689*
Fourth Service Command	60 6	4.5	18.8	16.1	100.0	*13,148*
Camp Blanding, Fla.	69.3	1.2	17.8	11.7	100.0	*1,180*
Camp Shelby, Miss.	53.2	6.4	22.1	18.3	100.0	*1,487*
Fort Benning, Ga.	70.7	—	17.2	12.1	100.0	*470*
Fort Bragg, N.C.	64.0	2.5	24.5	9.0	100.0	*1,817*
Fort Jackson, S.C.	55.6	12.7	14.0	17.7	100.0	*2,548*
Fort MacPherson, Ga.	64.0	3.2	14.8	18.0	100.0	*1,211*
Fort McClellan, Ala.	63.5	1.9	22.1	12.5	100.0	*1,431*
Fort Oglethorpe, Ga.	59.3	1.7	18.3	20.7	100.0	*3,004*
Fifth Service Command	62.2	0.5	23.5	13.8	100.0	*13,356*
Cleveland, Ohio	63.4	0.1	21.4	15.1	100.0	*2,753*
Columbus (Fort Hays), Ohio	66.7	0.1	19.8	13.4	100.0	*2,720*
Huntington, W. Va.	66.3	2.3	22.4	9.0	100.0	*2,819*
Indianapolis, Ind.	52.6	—	33.5	13.9	100.0	*2,592*
Louisville, Ky.	61.7	—	21.0	17.3	100.0	*2,472*

TABLE 2 (Continued)

SERVICE COMMAND AND INDUCTION STATION	DISPOSITION IN PERCENTAGES: Accepted	Rejected: Administrative	Rejected: Medical and physical	Rejected: Psychiatric	TOTAL: Per cent	TOTAL: Number
Sixth Service Command	63.5	0.9	21.8	13.8	100.0	*10,606*
Chicago, Ill.	71.2	1.8	19.4	7.6	100.0	*4,523*
Detroit, Mich.	52.6	*	25.8	21.6	100.0	*4,235*
Milwaukee, Wis.	69 8	0.8	18.3	11.1	100.0	*1,848*
Seventh Service Command	55.0	*	28.0	17.0	100.0	*11,780*
Fort Leavenworth, Kans.	53.9	—	30.4	15.7	100.0	*4,348*
Fort Logan, Colo.	54.9	0 3	28.6	16.2	100.0	*1,220*
Fort Snelling, Minn.	66.5	—	19.4	14.1	100.0	*3,129*
Jefferson Barracks, Mo.	44.7	*	33.4	21.9	100.0	*3,083*
Eighth Service Command	65.7	0.4	23.4	10.5	100.0	*8,321*
Camp Joseph T. Robinson, Ark.	65.2	0.1	26.5	8.2	100.0	*1,039*
Dallas, Texas	70.2	0.2	23.5	6.1	100.0	*1,458*
Fort Bliss, Tex.	67.9	2.7	23.9	5.5	100.0	*402*
Fort Sam Houston, Tex.	62.8	0.6	24.0	12.6	100.0	*943*
Houston, Tex.	69.1	0 2	22.5	8.2	100.0	*1,056*
New Orleans, La.	50.0	1.3	27.8	20.9	100.0	*863*
Oklahoma City, Okla.	70.2	*	19.4	10.4	100.0	*1,888*
Santa Fe, N.M.	66.4	—	19.5	14.1	100.0	*241*
Shreveport (Fort Hamburg), La.	58.5	—	27.6	13.9	100.0	*431*
Ninth Service Command	55.1	2.6	34.1	8.2	100.0	*12,387*
Butte, Mont.; Boise, Ida.; Spokane, Wash.	50.9	0.9	42.8	5.4	100.0	*1,268*
Camp Beale, Calif.	95.3	1.9	2.3	0.5	100.0	*214*
Fort Douglas, Utah	55.0	2.4	36.6	6.0	100.0	*816*
Fort Lewis, Wash.	73.5	9.5	13.2	3.8	100.0	*53*
Fort MacArthur, Calif.	70.5	2.8	16.2	10.5	100.0	*142*
Los Angeles, Calif.	56.8	1.7	34.1	7.4	100.0	*3,023*
Phoenix, Ariz.	47.6	0.2	45.6	6.6	100.0	*672*
San Francisco, Calif.	55.4	6.5	32.7	5.4	100.0	*3,226*
Seattle, Wash.; Portland, Oreg.	52.3	0.2	32.0	15.5	100.0	*2,973*

* Less than 0.05%.

drawn into these induction stations. This argument would be easier to support if the stations *within* a given region had somewhat the same rates and if the variability within regions was much less than the variability between regions.[4] But when Pittsburgh had

[4] Actually, the F-test shows that the variability was somewhat greater between regions than within regions, with results significant only at the .05 level.

TABLE 3

PSYCHIATRIC DIAGNOSES OF MEN REJECTED FOR PSYCHIATRIC REASONS, CLASSIFIED BY SERVICE COMMAND AND INDUCTION STATION
(August 1945)

SERVICE COMMAND AND INDUCTION STATION	PSYCHIATRIC DIAGNOSIS IN PERCENTAGES				TOTAL	
	Psycl neurc	*Psycl*	*Psycl*	*Other*	*Per cent*	*Number*
Total	39.9	27.8	2.1	30.2	100.0	*15,101*
First Service Command	57.2	23.1	5.1	14.6	100.0	*1,218*
Boston (Fort Banks), Mass.	57.9	19.4	2.5	20.2	100.0	*396*
Portland (Fort Preble), Me.	90.2	—	3.3	6.5	100.0	*92*
Manchester (Grenier Field), N.H.	76.8	2.4	13.0	7.8	100.0	*337*
New Haven, Conn.	25.0	57.9	1.1	16.0	100.0	*276*
Rutland, Vt.	52.9	23.5	1.5	22.1	100.0	*68*
Springfield, Mass.	40.8	42.9	2.0	14.3	100.0	*49*
Second Service Command	54.9	31.9	2.8	10.4	100.0	*1,835*
Albany-Syracuse, N.Y.	63.0	22.9	0.7	13.4	100.0	*135*
Buffalo-Rochester, N.Y.	33.0	35.3	0.6	31.1	100.0	*167*
Newark-Camden, N.J.	49.0	32.2	3.6	15.2	100.0	*251*
New York, N.Y	58.1	32.3	3.1	6.5	100.0	*1,282*
Third Service Command	18.4	32.6	1.3	47.7	100.0	*2,742*
Baltimore, Md.	8.1	8.4	0.7	82.8	100.0	*611*
Harrisburg, Pa.	17.2	52.0	1.2	29.6	100.0	*344*
Philadelphia, Pa.	2.7	17.1	7.8	72.4	100.0	*258*
Pittsburgh, Pa.	29.8	42.5	0.3	27.4	100.0	*1,113*
Richmond, Va.	21.2	54.6	4.5	19.7	100.0	*66*
Roanoke, Va.	14.3	54.6	1.0	30.1	100.0	*196*
Wilkes-Barre, Pa.	9.7	2.0	—	88.3	100.0	*154*
Fourth Service Command	37.3	28.8	1.2	32.7	100.0	*2,103*
Camp Blanding, Fla.	3.0	17.5	—	79.5	100.0	*137*
Camp Shelby, Miss.	16.9	81.3	—	1.8	100.0	*272*
Fort Benning, Ga.	28.1	68.3	—	3.6	100.0	*57*
Fort Bragg, N.C.	49.8	32.1	13.9	4.2	100.0	*165*
Fort Jackson, S.C.	14.3	32.3	0.2	53.2	100.0	*455*
Fort MacPherson, Ga.	48.6	15.2	1.0	35.2	100.0	*216*
Fort McClellan, Ala.	2.8	2.3	—	94.9	100.0	*178*
Fort Oglethorpe, Ga.	73.9	13.5	—	12.6	100.0	*623*
Fifth Service Command	35.8	14.0	0.8	49.4	100.0	*1,827*
Cleveland, Ohio	26.0	4.6	0.7	68.7	100.0	*416*
Columbus (Fort Hays), Ohio	32.8	14.2	0.3	52.7	100.0	*366*
Huntington, W. Va.	34.7	22.0	1.2	42.1	100.0	*254*
Indianapolis, Ind.	56.8	7.2	0.3	35.7	100.0	*361*
Louisville, Ky.	30.7	24.2	1.9	43.2	100.0	*430*

TABLE 3 (Continued)

SERVICE COMMAND AND INDUCTION STATION	PSYCHIATRIC DIAGNOSIS IN PERCENTAGES				TOTAL	
	Psycho-neurotic	*Psychopathic*	*Psychotic*	*Other*	*Per cent*	*Number*
Sixth Service Command	42.7	30.2	0.6	26.5	100.0	*1,465*
Chicago, Ill.	71.8	10.7	1.2	16.3	100.0	*344*
Detroit, Mich.	31.5	39.0	—	29.5	100.0	*915*
Milwaukee, Wis.	43.1	23.8	2.5	30.6	100.0	*206*
Seventh Service Command	38.5	30.5	3.1	27.9	100.0	*2,003*
Fort Leavenworth, Kans.	26.6	55.4	1.1	16.9	100.0	*684*
Fort Logan, Colo.	37.4	17.1	26.8	18.7	100.0	*198*
Fort Snelling, Minn.	43.0	15.9	0.4	40.7	100.0	*442*
Jefferson Barracks, Mo.	47.6	19.2	—	33.2	100.0	*679*
Eighth Service Command	63.8	24.5	0.9	10.8	100.0	*872*
Camp Joseph T. Robinson, Ark.	75.3	17.6	3.6	3.5	100.0	*85*
Dallas, Tex.	86.5	13.5	—	—	100.0	*89*
Fort Bliss, Tex.	54.5	45.5	—	—	100.0	*22*
Fort Sam Houston, Tex.	82.3	6.8	0.8	10.1	100.0	*119*
Houston, Tex.	62.1	18.4	2.2	17.3	100.0	*87*
New Orleans, La.	74.6	2.2	—	23.2	100.0	*181*
Oklahoma City, Okla.	22.7	70.9	—	6.4	100.0	*195*
Santa Fe, N.M.	88.3	—	—	11.7	100.0	*34*
Shreveport (Fort Hamburg), La.	61.6	25.0	3.4	10.0	100.0	*60*
Ninth Service Command	40.1	31.2	4.2	24.5	100.0	*1,036*
Butte, Mont.; Boise, Ida.; Spokane, Wash.	70.5	11.8	2.9	14.8	100.0	*68*
Camp Beale, Calif.	100.0	—	—	—	100.0	*1*
Fort Douglas, Utah	42.9	40.8	—	16.3	100.0	*49*
Fort Lewis, Wash.	—	100.0	—	—	100.0	*2*
Fort MacArthur, Calif.	6.7	86.6	6.7	—	100.0	*15*
Los Angeles, Calif.	46.8	27.0	1.4	24.8	100.0	*222*
Phoenix, Ariz.	35.6	42.2	2.2	20.0	100.0	*45*
San Francisco, Calif.	69.1	20.4	0.6	9.9	100.0	*171*
Seattle, Wash.; Portland, Oreg.	22.9	35.9	7.8	33.4	100.0	*463*

3 times the proportion of psychiatric rejects of Philadelphia, when Detroit had 3 times the proportion of Chicago, New Orleans 3 times the proportion of Dallas, and Seattle-Portland 3 times the proportion of San Francisco, it is difficult to believe that the standards were the same in all places. Similarly, it is hard to believe that the same standards of diagnosis were being used when, among psychiatric rejects, New Haven found 3 times as large a proportion of psychopaths as Boston, Pittsburgh found 5 times as large a proportion as

Baltimore, or the Camp Shelby, Mississippi, station found 35 times as large a proportion as the Fort McClellan, Alabama, station.

In the absence of a yardstick providing a common nationwide standard, it has not been possible in the past to do more than express doubts as to the uniformity of diagnoses. However, for the more than one hundred thousand men represented in Tables 2 and 3 we do have a standard against which this variability can be judged. This is the NSA. Let us look first at the variability of the proportions screened among the different induction stations and see how the NSA findings compare with the actual psychiatric diagnoses.

NSA Scores as Compared with Psychiatric Diagnoses

Table 4 gives us the basic information. The cutting point on the NSA scores, plus the additional information from "critical signs," is so adjusted that, the nation over, 31.1 per cent of all men were screened. By screened we do not mean, as the previous chapter should have made clear, that the test diagnosed a given man as a psychiatric case. Rather it simply meant that here was a man to whom the busy examining physician should give much more than average attention. In the group of men screened by the test should be found most of the men eventually diagnosed as psychiatric cases, and, because of fallibility both of the test and of the diagnoses, there would also be found a good many men who would eventually be labeled as fit for military service. If, however, the test screened twice as many men at one induction station as at another, this should indicate a probability that the psychiatric population was higher at the former station than at the latter.

Actually, as Table 4 shows, there was relatively small variability from station to station in the percentages screened by the NSA. Whereas the psychiatric rejection rate, as we have seen in Table 2, ranged from 0.5 per cent to 50.6 per cent, the data in Table 4 show that the percentages screened by the NSA ranged only from 21.2 per cent to 44.5 per cent, and the standard deviation was only 5.1 per cent.

Moreover—and this is a very important point—the variability in NSA scores was decidedly due more to the differences between regions than to differences between stations within the same region.[5]

[5] The F-test shows that the tendency for homogeneity within service commands as compared with the variability between commands is significant at the .01 level.

TABLE 4

PROPORTION SCREENED BY NSA AMONG PREINDUCTION EXAMINEES, CLASSIFIED BY DISPOSITION BY SERVICE COMMAND AND INDUCTION STATION (August 1945)

SERVICE COMMAND AND INDUCTION STATION	PERCENTAGES RECEIVING CRITICAL SIGNS AND/OR CRITICAL SCORES AMONG:							
			REJECTIONS*					
				Psychiatric				
	All men examined	*Acceptable men*	*Medical and physical*	*Total*	*Psychoneurotic*	*Psychopathic*	*Psychotic*	*Other*
Total	31.1	21.8	30.3	69.5	80.8	68.2	70.8	56.4
First Service Command	28.1	14.3	25.0	67.2	74.2	57.0	66.0	58.6
Boston (Fort Banks), Mass.	26.8	14.4	28.1	68.6	77.3	53.3	80.0	54.8
Portland (Fort Preble), Me.	22.6	13.2	17.9	58.9	58.8	a	66.7	50.0
Manchester (Grenier Field), N.H.	44.5	12.6	18.8	73.6	75.4	100.0	61.5	69.0
New Haven, Conn.	26.2	14.0	27.4	65.7	78.4	60.7	100.0	61.5
Rutland, Vt.	24.9	16.5	17.2	58.7	72.4	37.5	0.0	53.3
Springfield, Mass.	21 2	15.4	20.5	59.0	80.0	38.3	100.0	57.1
Second Service Command	30.4	22.5	28.0	72.0	78.7	68.5	80.3	44.8
Albany-Syracuse, N.Y.	27.2	18.2	29.5	74.0	79.1	81.7	100.0	39.1
Buffalo-Rochester, N.Y.	30.5	24 0	30.4	69.0	77.6	66.2	0.0	63.5
Newark-Camden, N.J.	27.1	20.8	27.2	70.1	78.7	71.7	88.9	34.0
New York, N.Y.	31.9	23.8	26.8	72.4	78.5	67.3	80.0	38.4
Third Service Command	32.7	22.6	31.0	69.1	78.1	69.5	88.9	65.0
Baltimore, Md.	37.4	27.8	29.7	74.4	90.0	74.6	75.0	72.7
Harrisburg, Pa	35.0	20.5	32.8	76.6	89.9	71.1	100.0	77.4
Philadelphia, Pa.	32.1	23.6	33.3	83.4	85.7	84.0	100.0	81.4
Pittsburgh, Pa.	32.5	19.9	28.3	61.0	73.0	64.2	66.7	43.4
Richmond, Va.	31.6	25.0	41.4	74.1	78.6	80.7	33.3	61.5
Roanoke, Va.	35.7	24.0	37.2	66.3	78.6	79.4	100.0	37.4
Wilkes-Barre, Pa.	24.0	16.6	23.3	69.7	93.3	66.7	a	66.8
Fourth Service Command	36.1	26.1	37.3	76.6	86.0	82.1	100.0	59.5
Camp Blanding, Fla.	36.2	28.6	35.8	79.5	75.0	75.0	a	80.7
Camp Shelby, Miss.	42.4	25.1	41 9	84.0	93.4	82.6	a	60.0
Fort Benning, Ga.	36.4	28.8	39.3	77.2	93.7	71.8	a	50.0
Fort Bragg, N.C.	39.9	28.9	45.7	95.8	92.8	98.1	100.0	100.0
Fort Jackson, S.C.	30.0	25.2	29.9	57.5	95.5	85.1	100.0	30.5
Fort MacPherson, Ga.	34.6	24.7	35.5	70.4	82.7	63.4	100.0	55.0
Fort McClellan, Ala.	37.5	26.9	38.4	93.2	100.0	100.0	a	92.8
Fort Oglethorpe, Ga.	36.6	23.1	33.4	78.8	83.3	79.7	a	50.6

* Administrative rejections have been omitted because of the small number of men in this category.

a No men in this category.

TABLE 4 (Continued)

SERVICE COMMAND AND INDUCTION STATION	PERCENTAGES RECEIVING CRITICAL SIGNS AND/OR CRITICAL SCORES AMONG:								
			REJECTIONS						
				Psychiatric					
	All men examined	*Acceptable men*	*Medical and physical*	*Total*	*Psychoneurotic*	*Psychopathic*	*Psychotic*	*Other*
Fifth Service Command	26.7	17.7	27.7	66.8	87.9	68.1	81.1	51.2
Cleveland, Ohio	29.2	19.2	32.8	70.1	98.2	79.0	66.7	58.9
Columbus (Fort Hays), Ohio	25.4	19.0	28.0	57.4	83.3	73.0	100.0	35.7
Huntington, W. Va	22.2	16.5	23.3	68.8	86.4	69.9	33.3	53.1
Indianapolis, Ind.	27.6	16.9	25.5	73.1	88.7	64.4	100.0	48.7
Louisville, Ky.	28.7	18.1	30.1	67.1	83.3	63.6	100.0	55.8
Sixth Service Command	27.3	19.6	26.3	64.6	80.9	59.7	77.8	42.9
Chicago, Ill.	24.1	19.5	25.9	64.6	69.0	62.2	75.0	42.9
Detroit, Mich.	26.9	15.3	22.2	60.9	88.8	58.2	a	35.3
Milwaukee, Wis.	35.1	26.6	39.7	78.9	86.5	69.3	80.0	74.7
Seventh Service Command	29.7	18.9	29.4	64.4	77.6	65.0	49.4	47.9
Fort Leavenworth, Kans.	32.1	22.7	34.2	63.2	77.5	65.8	50.0	33.7
Fort Logan, Colo.	30.1	18.4	31.9	66.1	83.8	55.9	49.2	64.8
Fort Snelling, Minn.	24.1	14.6	27.3	64.3	79.9	51.5	50.0	52.6
Jefferson Barracks, Mo.	31.5	19.2	24.5	66.0	74.8	73.8	a	48.8
Eighth Service Command	33.9	27.1	33.6	77.1	83.5	65.4	75.0	66.1
Camp Joseph T. Robinson, Ark.	32.3	24.3	36.6	70.9	78.2	53.6	66.7	0.0
Dallas, Tex.	33.8	29.7	39.2	70.0	68.9	74.7	a	a
Fort Bliss, Tex.	33.8	28.8	36.2	81.9	91.7	70.0	a	a
Fort Sam Houston, Tex.	26.9	16.7	25.7	81.5	85.7	62.5	0.0	66.7
Houston, Tex.	36.3	29.9	34.2	87.2	80.8	68.8	100.0	93.3
New Orleans, La.	34.3	20.2	26.8	77.9	83.9	75.0	a	59.4
Oklahoma City, Okla.	37.3	32.4	35.1	76.0	96.0	66.7	a	100.0
Santa Fe, N.M.	33.6	24.2	36.1	76.5	83.3	a	a	25.0
Shreveport (Fort Hamburg), La.	33.5	25.6	29.7	75.0	89.2	53.6	100.0	33.3
Ninth Service Command	32.9	25.0	35.5	72.6	84.2	65.8	59.3	65.3
Butte, Mont.; Boise, Ida.; Spokane, Wash.	32.8	23.5	38.0	81.0	85.0	75.0	100.0	80.0
Camp Beale, Calif.	30.2	28.8	60.0	0.0	0.0	a	a	a
Fort Douglas, Utah	28.9	22.5	31.4	67.0	76.2	65.0	a	50.0
Fort Lewis, Wash.	22.8	12.9	42.9	100.0	a	100.0	a	a
Fort MacArthur, Calif.	25.9	16.0	30.1	86.6	100.0	84.6	100.0	a
Los Angeles, Calif.	32.9	24.1	36.4	80.0	87.4	76.6	100.0	68.9
Phoenix, Ariz.	29.3	16.8	35.6	73.6	81.1	63.4	100.0	77.8
San Francisco, Calif.	35.3	29.2	39.4	75.8	80.5	66.1	0.0	64.6
Seattle, Wash.; Portland, Oreg.	33.2	24.8	30.6	66.1	87.9	59.4	53.2	63.4

The largest proportions screened by the NSA were in the South, the smallest proportions in the Middle West and New England.

Now, these two sets of facts, (1) that psychiatric diagnosis varied widely from station to station and (2) that NSA scores had much less variation,[6] still may have little significance unless it can be shown that, in spite of these overall differences, the men screened by the NSA did tend to include a majority of the psychiatric rejects.

What are the facts? Let us look again at Table 4. The top line of this table summarizes the findings for the country as a whole. It shows the following proportions screened by the NSA:

Among all men examined	31.1%
Among all men accepted for service	21.8
Among nonpsychiatric rejects	30.3
Among all psychiatric rejects	69.5

If we break down the psychiatric rejects further, we find that the test was particularly effective in screening men diagnosed as psychoneurotic, less so in screening other classes of psychiatric rejects. Proportions screened:

Among psychoneurotic rejects	80.8%
Among psychopathic rejects	68.2
Among psychotic rejects	70.8
Among all other psychiatric rejects	56.4

As we see, the NSA, in spite of the variability in psychiatric diagnosis, was successful in picking out four out of five of the men subsequently diagnosed as psychoneurotic, at a cost of also screening 21.8 per cent of the men subsequently passed as fit for military service. This is about the same order of discrimination which we saw in the earlier results reported in the previous chapter. But the test was, as had been feared in advance, not quite so successful in spotting the kind of people diagnosed as psychopathic, screening only 68.2 per cent. It was even less successful with the "all other" category, which included a great variety of things and which was highly varied from one induction station to another.[7]

We must repeat that the test and the clinical diagnosis were not necessarily experimentally independent. But these results do show

[6] The coefficient of variation of psychiatric rejection rates is five times that of proportions screened by the NSA.

[7] At one station, a representative from the Research Branch found physicians rejecting as "emotionally immature" almost all boys under twenty who were not yet sprouting a respectable beard.

that, whatever the test was measuring, it was measuring verbal behavior which tended to discriminate between men accepted for service and those rejected on psychiatric grounds. The significance of this finding is strengthened by the further finding, shown in Table 4, that at every individual induction station with at least fifty psychiatric rejects the NSA successfully screened a majority of those rejected. Also, at individual stations throughout the country, the tendency was for the test to be most effective in detecting psychoneurotics, somewhat less so in detecting other categories. Among 48 induction stations for which complete diagnostic comparisons are possible,[8] psychoneurotics were detected by the NSA more successfully than either psychopaths or "others" in 40 cases, and psychopaths were more frequently screened by the NSA than "others" in 31 cases. The completely consistent pattern of most success with psychoneurotics, next with psychopaths, and least with others is observable in 27 out of the 48 cases. This suggests, at least, that in spite of the wide variations from station to station in labels pinned on the psychiatric rejects, the particular diagnoses were not wholly semantic accidents, depending merely on the personal point of view of the local doctors. If *exactly* the same kind of men had been called psychoneurotics at one station and psychopaths at another, we hardly would have observed the uniformity with which the NSA screened psychoneurotics more successfully than psychopaths. There must have been at least a central core of agreement on diagnosis from station to station, as well as, to be sure, great divergency.[9]

The variations between induction stations, in proportions of all men accepted after medical examination, in psychiatric rejection rates, in the proportions of all men screened by the NSA, and in the proportions of psychiatric rejects screened by the NSA, may be seen readily in Table 5, which puts into summary form much of the data we have been discussing from Tables 2, 3, and 4.

We now have seen (1) that psychiatric rates varied widely, (2)

---

[8] Omitting psychotics because of the small number of cases.

[9] To put it another way, if we assume a fairly fixed proportion of psychopaths in each population of psychiatric rejects and ascribe deviations from the proportion to mere semantical differences, then the larger the proportion classified psychopathic the larger the proportion of actual psychoneurotics that would be included among them. Since psychoneurotics were more successfully detected by the NSA, it would have to follow that the larger the proportion of psychiatric rejections classified as psychopaths, the larger the proportion of psychopaths that would be detected by the NSA. This is not the case however. In each instance, the proportion of a diagnostic group detected by the NSA was independent of the proportion that group was of the psychiatric rejections.

that the proportions screened by NSA varied much less widely, (3) that for the country as a whole and for individual stations the test did discriminate, especially with respect to psychoneurotics. Yet the fact is that the small variation in the proportions screened by the NSA in hardly any way accounts for the large variation

TABLE 5

SUMMARY OF INTERSTATION VARIATIONS IN REJECTION RATES AND PROPORTIONS SCREENED BY NSA

	FREQUENCY OF GIVEN RATE BY SERVICE COMMAND									
	1st	*2nd*	*3rd*	*4th*	*5th*	*6th*	*7th*	*8th*	*9th*	*Total*
Proportion of all men accepted										
Under 50.0	1	1	1				1		1	5
50.0–53.9			2	1	1	1	1	1	2	9
54.0–57.9	1		3	1			1		3	9
58.0–61.9	2	2		1	1			1		7
62.0–65.9	2			3	1			2		8
66.0–69.9		1		1	2	1	1	3		9
Over 69.9			1	1		1		2	3	8
Psychiatric rejection rate										
Under 6.0			1					1	4	6
6.0– 8.9			1			1		3	3	8
9.0–11.9	1	3	1	2	1	1		1	1	11
12.0–14.9		1		2	2		1	3		9
15.0–17.9	2			1	2		2		1	8
18.0–20.9	2		2	3				1		8
Over 20.9	1		2			1	1			5
Proportion of all men screened by NSA										
Under 24.0	2				1				1	4
24.0–26.9	3		1		1	2	1	1	1	10
27.0–29.9		2			3				2	7
30.0–32.9		2	3	1			3	1	3	13
33.0–35.9			2	1		1		5	2	11
36.0–38.9			1	4				2		7
Over 38.9	1			2						3
Proportion of psychiatric rejects screened by NSA										
Under 60.0	3			1	1				1	6
60.0–64.9			1			2	2			5
65.0–69.9	2	1	2		2		2		2	11
70.0–74.9	1	3	2	1	2			2	1	12
75.0–79.9			1	3		1		4	1	10
80.0–84.9			1	1				2	2	6
Over 84.9				2				1	2	5
Number of induction stations	6	4	7	8	5	3	4	9	9	55

in psychiatric rejection rates among individual stations. If we exclude the two extreme stations in psychiatric rejection rates (Manchester, New Hampshire, and Camp Beale, California), the correlation coefficient for fifty-three stations between the two variables is only .11 ± .14.

To trace further the reasons for the wide differences from one service command to another and from one induction station to another in psychiatric rejection rates would lead us into realms of speculation for which there are little supporting data. It should be noted, however, that the fact that the two service commands covering most of the South were highest in the proportion screened by the NSA is quite consistent with the findings in the Army, reported in Volume II, Chapter 8, which showed that men of lowest educational level had the highest psychiatric rates—both as measured by the Research Branch test scores and as demonstrated in actual practice in the Army. Since the South has much the lowest educational level in the country, it would be expected that it would have a somewhat higher proportion of preinductees screened by the NSA than other regions in the country. But among the nine service commands, the two covering most of the South were not the highest in psychiatric rejection rates. The Eighth Service Command in the Southwest had a psychiatric rejection rate of 10.5 per cent, one of the lowest in the country, while three other service commands had higher rejection rates than the Fourth, which is in the Southeast. In so far as there were regional differences in rejection rates, these differences may be more closely related to some prevailing administrative policy than to any other factor. About all that can be said for the variations in diagnosis and disposition *within* Service Commands is that there is a strong likelihood that they reflect different standards used by the doctors at different stations.

Further Observations on the Value of Critical Signs to Supplement NSA Scores

Chart I shows, for each numerical score on the NSA, the disposition of examinees in August 1945. It will be seen that the percentage rejected as psychiatric dropped fairly regularly, tending to flatten out after a score of 22 was reached.

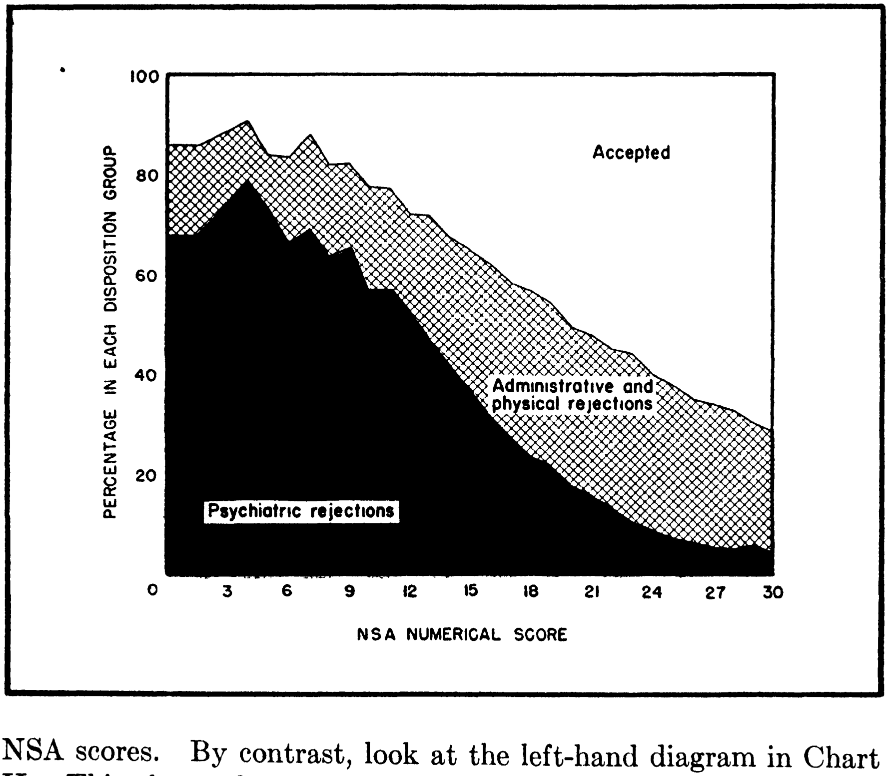
In Chart I, the data are shown for all men, whether or not they showed up as possessing "critical signs"—that is, gave an unfavor-

able response to any of the supplementary questions described in the previous chapter.

The efficacy of considering as screened all men with one or more critical signs, irrespective of their NSA score, is shown in Chart II. On the right is shown the disposition of cases, by NSA score, among those who did not receive critical signs. There the proportion of psychiatric rejects drops off very sharply among men making high

CHART I

Disposition of Preinduction Examinees,
Classified by NSA Numerical Score
(August 1945)



NSA scores. By contrast, look at the left-hand diagram in Chart II. This shows the disposition by NSA scores among those with one or more critical signs. There is still a relationship between the proportion of psychiatric rejects and test scores, but *even among those with high NSA scores* about 30 per cent, on the average, were psychiatric rejects.

Who were the men who had one or more critical signs, but made high NSA scores? We can see in Chart III. Here the psychiatric

CHART II

Disposition of Preinduction Examinees, Classified by NSA Numerical Score and Presence or Absence of Critical Signs
(August 1945)

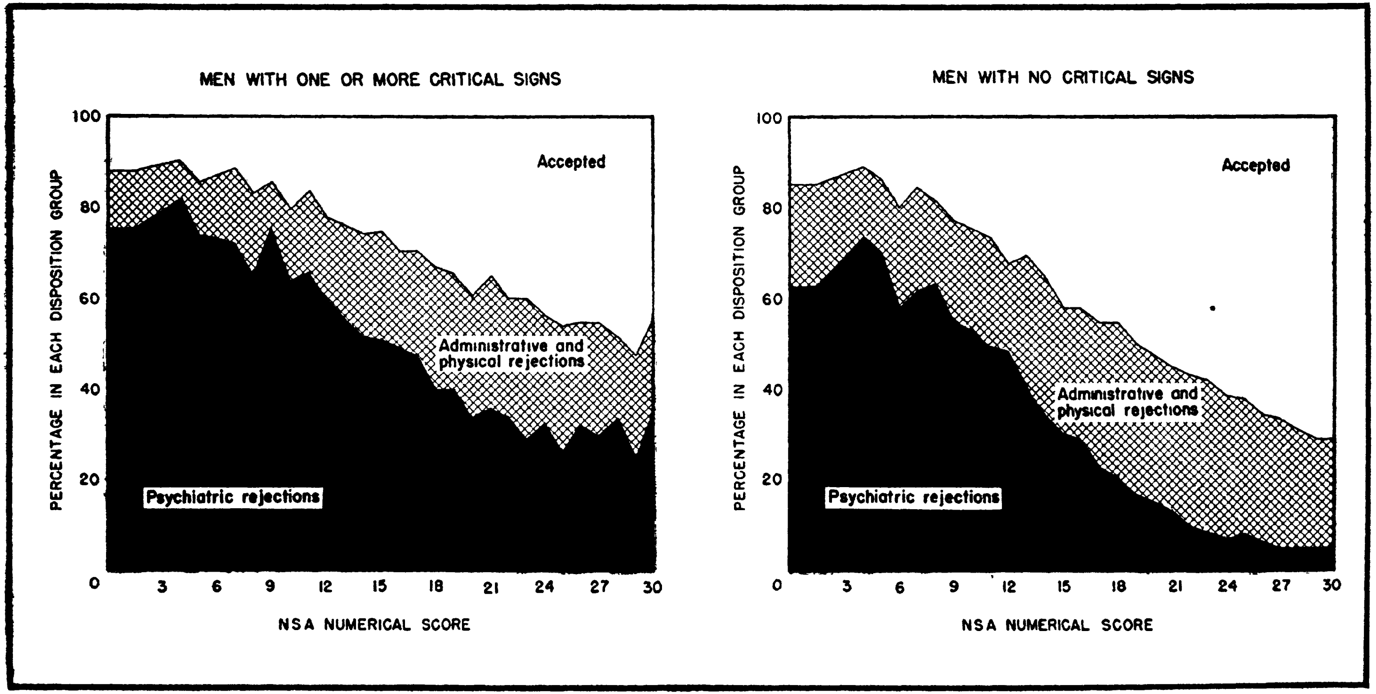
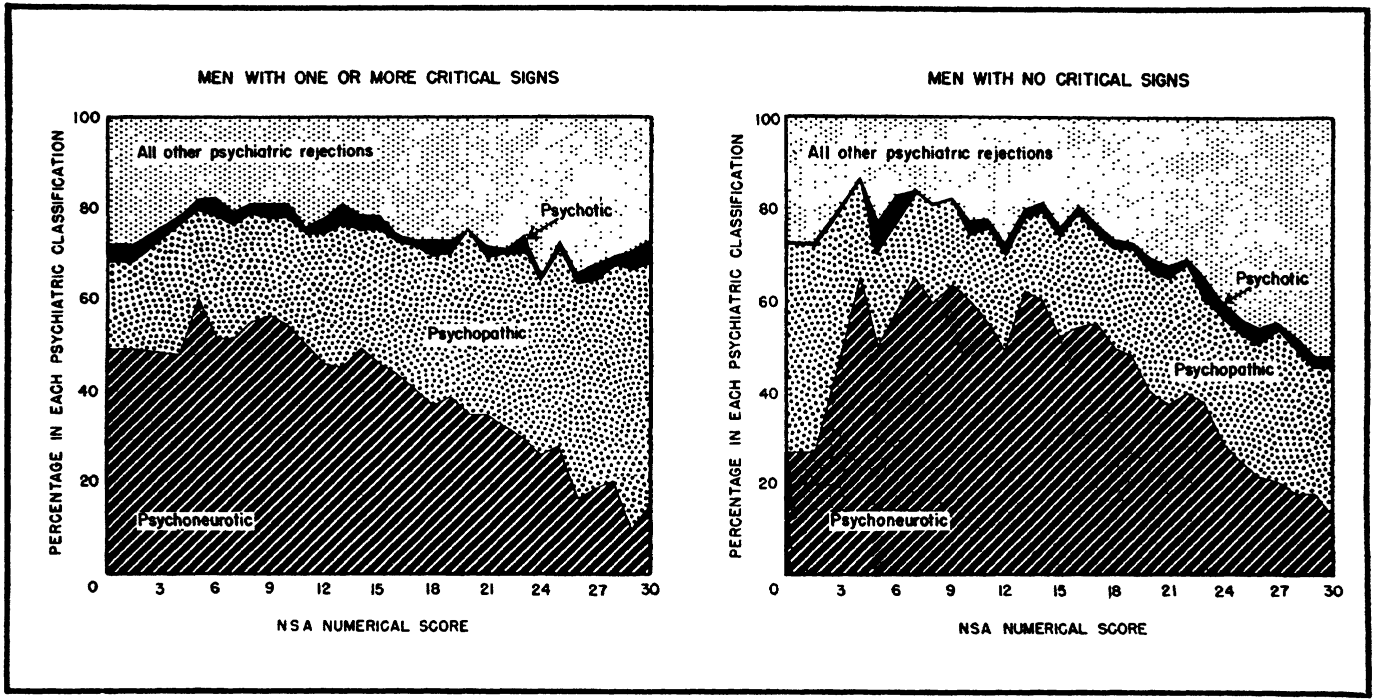


CHART III

DIAGNOSES OF PSYCHIATRIC REJECTIONS, CLASSIFIED BY NSA NUMERICAL SCORE AND PRESENCE OR ABSENCE OF CRITICAL SIGNS
(August 1945)



rejects with one or more critical signs are represented as 100 per cent for each NSA score interval. It becomes evident at once that as the NSA score increased the proportion of psychiatric rejections screened as *psychoneurotic went down*, while the proportion screened as *psychopathic went up*. The proportion of psychotics and "other psychiatric cases" to all psychiatric cases was relatively constant. The addition of critical signs to the original test did serve its intended purpose to some extent, in that they tended to screen a substantial number of psychopaths, as well as others not psychoneurotic, who otherwise would have made acceptable scores.

Nevertheless, as we have seen, so varying was the psychiatric diagnosis from station to station, even the use of signs did not enable the test to discriminate what the psychiatrists would do about psychopaths and others as well as it discriminated what they would do about psychoneurotics.

For the benefit of the student who wishes to study more in detail the effectiveness of the test with or without signs, for various diagnoses, Table 6 is presented, summarizing the NSA scores, for all types of men found acceptable or rejected.

It can be determined from the data in Table 6 that the most efficient cutting point for discriminating psychiatric rejections from acceptable men, on the basis of their NSA scores alone, would appear to be about 21, which is a little higher than the cutting point of 19 when the same test was given to Army recruits after three months of service. At the cutting point of 21, the percentages screened by either sign or score[10] were 21.8 per cent among men found acceptable for service and 69.5 per cent among men rejected for psychiatric reasons. While this represents a rather high level of differentiation, it does not reach the level of detection to be desired if the test were to be used for *eliminating* individual psychiatric examinations in the majority of cases. At this level of performance, almost a third of the men psychiatrists rejected would not even have come to their attention. The cutting point of the test could be shifted to a point at which the detection of psychiatric rejections would reach an acceptable level, say a score of 26 used in conjunction with critical signs, at which point 87.2 per cent of the psychiatric rejections were screened. But such a shift would make the test not too efficient at the other end, requiring the examination of 57.7 per cent of the men found to be acceptable. This would have saved the

[10] Here, as before, the actual cutting point is itself a critical score, so that we screen all men possessing a critical sign or having a score of 21 or less.

TABLE 6

Percentage Distribution of NSA Scores of Preinduction Examinees, According to Their Classification (August 1945)

Score	All Men Examined: Frequency of Given Score			Men Acceptable for Service: Frequency of Given Score			Rejections: Administrative: Frequency of Given Score			Rejections: Medical and Physical: Frequency of Given Score			Rejections: Psychiatric**: All Psychiatric: Frequency of Given Score			Rejections: Psychiatric**: Psychoneurotic: Frequency of Given Score			Rejections: Psychiatric**: Psychopathic: Frequency of Given Score			Rejections: Psychiatric**: Psychotic: Frequency of Given Score			Rejections: Psychiatric**: Other: Frequency of Given Score		
	Total	*Without sign*	*With sign*	*Total*	*Without sign*	*With sign*	*Total*	*Without sign*	*With sign*	*Total*	*Without sign*	*With sign*	*Total*	*Without sign*	*With sign*	*Total*	*Without sign*	*With sign*	*Total*	*Without sign*	*With sign*	*Total*	*Without sign*	*With sign*	*Total*	*Without sign*	*With sign*
30	7.5	7.3	0.2	9.1	9.0	0.1	12.5	11.9	0.6	6.0	5.9	0.1	3.0	2.6	0.4	1.1	0.9	0.2	4.0	3.2	0.8	3.2	1.9	1.3	4.7	4.3	0.4
29	8.8	8.5	0.3	10.8	10.5	0.3	7.9	7.2	0.7	7.4	7.2	0.2	3.5	3.0	0.5	1.5	1.4	0.1	4.2	3.2	1.0	4.5	3.2	1.3	5.4	4.9	0.5
28	10.2	9.8	0.4	11.9	11.6	0.3	9.2	8.5	0.7	9.7	9.4	0.3	4.5	3.6	0.9	2.1	1.6	0.5	5.7	4.2	1.5	3.2	1.9	1.3	6.6	5.7	0.9
27	10.1	9.7	0.4	11.6	11.2	0.4	9.4	9.0	0.4	10.2	9.8	0.4	4.5	3.6	0.9	2.3	1.9	0.4	5.8	4.3	1.5	3.8	1.9	1.9	6.2	5.3	0.9
26	9.4	8.9	0.5	10.5	10.1	0.4	10.9	10.2	0.7	9.4	9.0	0.4	4.8	3.6	1.2	2.5	2.0	0.5	5.7	3.7	2.0	5.1	4.5	0.6	7.0	5.6	1.4
25	8.3	7.8	0.5	9.0	8.5	0.5	7.7	6.8	0.9	9.0	8.5	0.5	4.8	3.8	1.0	3.1	2.4	0.7	5.1	3.5	1.6	5.4	4.1	1.3	6.6	5.7	0.9
24	7.3	6.7	0.6	7.5	7.1	0.4	7.3	6.8	0.5	8.2	7.7	0.5	4.8	3.5	1.3	3.4	2.5	0.9	5.3	3.6	1.7	5.4	4.1	1.3	6.1	4.6	1.5
23	6.2	5.6	0.6	6.0	5.6	0.4	5.2	4.6	0.6	7.4	6.8	0.6	4.6	3.4	1.2	4.0	3.1	0.9	4.6	2.8	1.8	7.9	5.4	2.5	5.1	4.1	1.0
22	5.2	4.6	0.6	5.0	4.6	0.4	4.3	3.5	0.8	6.0	5.4	0.6	4.8	3.4	1.4	4.5	3.4	1.1	5.2	3.3	1.9	3.8	2.2	1.6	4.7	3.4	1.3
21	4.3	3.7	0.6	3.9	3.5	0.4	3.3	2.7	0.6	4.9	4.3	0.6	4.9	3.3	1.6	4.6	3.1	1.5	5.2	3.3	1.9	6.0	3.8	2.2	5.1	3.6	1.5
20	3.6	3.0	0.6	3.2	2.8	0.4	4.3	3.2	1.1	4.0	3.4	0.6	4.6	3.2	1.4	4.6	3.4	1.2	5.0	3.0	2.0	3.2	3.2	0.0	4.4	3.3	1.1
19	3.0	2.4	0.6	2.5	2.1	0.4	2.8	2.4	0.4	3.4	2.8	0.6	4.7	2.9	1.8	5.3	3.6	1.7	4.4	2.5	1.9	4.5	1.3	3.2	4.1	2.6	1.5
18	2.7	2.1	0.6	2.0	1.7	0.3	3.1	2.1	1.0	3.0	2.5	0.5	4.5	2.9	1.6	5.1	3.6	1.5	4.2	2.4	1.8	5.7	1.6	4.1	3.9	2.5	1.4
17	2.3	1.8	0.5	1.7	1.4	0.3	2.3	1.8	0.5	2.6	2.1	0.5	4.6	2.8	1.8	5.7	3.9	1.8	4.1	2.1	2.0	3.5	1.6	1.9	3.7	2.1	1.6
16	1.9	1.4	0.5	1.3	1.0	0.3	1.9	1.4	0.5	1.8	1.4	0.4	4.6	2.7	1.9	5.3	3.6	1.7	4.4	2.4	2.0	3.2	1.6	1.6	3.3	1.7	1.6
15	1.7	1.2	0.5	1.0	0.8	0.2	1.8	0.9	0.9	1.6	1.2	0.4	4.3	2.5	1.8	5.3	3.2	2.1	3.9	1.9	2.0	4.7	2.2	2.5	3.3	2.0	1.3
14	1.5	1.0	0.5	0.8	0.6	0.2	1.3	1.1	0.2	1.3	0.9	0.4	4.3	2.4	1.9	6.1	3.7	2.4	3.5	1.7	1.8	4.7	2.2	2.5	2.8	1.5	1.3
13	1.3	0.8	0.5	0.6	0.4	0.2	1.1	0.5	0.6	1.1	0.8	0.3	3.9	2.1	1.8	5.5	3.4	2.1	3.4	1.3	2.1	3.8	1.3	2.5	2.5	1.4	1.1
12	1.1	0.6	0.5	0.5	0.3	0.2	0.5	0.2	0.3	0.8	0.5	0.3	3.9	2.0	1.9	4.7	2.4	2.3	3.4	1.5	1.9	5.6	2.8	2.8	3.2	1.8	1.4
11	0.9	0.5	0.4	0.4	0.3	0.1	1.1	0.5	0.6	0.6	0.4	0.2	3.7	1.7	2.0	4.9	2.4	2.5	3.1	1.4	1.7	1.3	0.3	1.0	2.9	1.3	1.6
10	0.8	0.4	0.4	0.3	0.2	0.1	0.4	0.2	0.2	0.5	0.3	0.2	3.0	1.4	1.6	4.3	2.1	2.2	2.1	0.8	1.3	3.2	1.3	1.9	2.0	1.0	1.0
9	0.6	0.3	0.3	0.2	0.1	0.1	0.5	0.2	0.3	0.3	0.2	0.1	2.6	1.0	1.6	4.1	1.7	2.4	1.9	0.7	1.2	1.3	0.0	1.3	1.6	0.6	1.0
8	0.5	0.2	0.3	0.2	0.1	0.1	0.1	0.0	0.1	0.3	0.1	0.2	2.1	0.9	1.2	3.1	1.4	1.7	1.7	0.7	1.0	0.6	0.0	0.6	1.3	0.6	0.7
7	0.3	0.1	0.2	*	*	*	0.4	0.2	0.2	0.2	0.1	0.1	1.5	0.5	1.0	2.2	0.9	1.3	1.2	0.3	0.9	1.3	0.0	1.3	1.0	0.3	0.7
6	0.2	0.1	0.1	*	*	*	0.1	0.0	0.1	0.2	0.1	0.1	1.2	0.4	0.8	1.6	0.6	1.0	1.0	0.3	0.7	2.6	1.0	1.6	0.6	0.2	0.4
5	0.2	0.1	0.1	*	*	*	0.0	0.0	0.0	0.1	*	0.1	0.9	0.3	0.6	1.2	0.3	0.9	0.6	0.2	0.4	1.2	0.6	0.6	0.6	0.2	0.4
4	0.1	*	0.1	*	*	*	0.0	0.0	0.0	*	*	*	0.7	0.2	0.5	1.0	0.4	0.6	0.7	0.2	0.5	0.3	0.0	0.3	0.5	0.1	0.4
3	*	*	*	*	*	*	0.1	0.0	0.1	*	*	*	0.4	0.2	0.2	0.4	0.1	0.3	0.4	0.2	0.2	1.0	0.0	1.0	0.4	0.2	0.2
2	*	*	*	*	*	*	0.0	0.0	0.0	*	*	*	0.2	*	0.2	0.3	*	0.3	0.1	*	0.1	0.0	0.0	0.0	0.3	0.1	0.2
1	*	*	*	*	*	*	0.0	0.0	0.0	*	*	*	0.1	*	0.1	0.2	*	0.2	0.1	*	0.1	0.0	0.0	0.0	*	0.0	*
0	*	*	*	*	0.0	*	0.5	0.5	0.0	*	*	0.0	*	*	*	*	*	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	*
Total %	100.0	88.6	11.4	100.0	93.5	6.5	100.0	86.4	13.6	100.0	90.8	9.2	100.0	63.9	36.1	100.0	63.0	37.0	100.0	58.7	41.3	100.0	54.0	46.0	100.0	70.8	29.2
Number	*107,385*			*61,881*			*1,249*			*29,154*			*15,101*			*6,010*			*4,213*			*315*			*4,563*		

* Less than 0.05%

** Includes rejections for both psychiatric and other medical reasons.

individual examination of only about one third of the men passing through the induction stations at this time. While even this result would represent some saving of psychiatrists' time, it fell short of the results originally expected.

Nevertheless, in view of the very brief time available to physicians for the psychiatric examination, there can be no doubt that the test score and critical signs should have served as a useful signal that here was a man who needed a thorough going over, while here was a man who could be given a cursory inspection. If this practice had been carried out in the earlier periods of the war before the NSA, when the induction stations were clogged and hundreds of thousands of men slipped by only to be caught later in the Army or Navy and discharged for psychiatric reasons, the saving in efficiency to the Army and in money to the American taxpayer might have been enormous. Until, however, psychiatric diagnosis is better standardized than it is today, it is likely that the predictability of psychiatric screening tests will fall far short of ideal.

CHAPTER 15

STUDIES OF THE POSTWAR PLANS OF SOLDIERS: A PROBLEM IN PREDICTION[1]

Introduction

As has been observed in Chapter 12, the present chapter and that following may be regarded as presenting the case history of a research project. Data on soldiers' plans were utilized to make predictions of the number of Army veterans who might be expected to enter certain fields of endeavor after the war. The predictions were not the result of refined statistical techniques. They had no basis in previously observed relationships between prediction items and a criterion. They involved a number of assumptions which could not be checked and a number of contingencies which could not be evaluated until after soldiers were discharged in large numbers. Nevertheless, they were widely used and proved on the whole to be reasonably accurate forecasts. Two follow-up studies to evaluate the extent to which soldiers carried out or attempted to carry out the plans they expressed prior to discharge permit an evaluation of the procedures used and give certain clues as to the conditions under which expressed intentions and plans are predictive of behavior.

Relatively early in the war, various groups expressed interest in finding out what effect Army service was likely to have on the occupational plans and intentions of the selectee. Employers were interested in estimating the number of their former employees likely to return to their old jobs; educators were interested in learning how many of the men might want to return to school. Many government departments whose programs included aid to returning veterans were seeking accurate information on the demands which

[1] By John A. Clausen. The initial study of soldiers' postwar plans was developed by Lyonel C. Florant and Louis Guttman. The main surveys in this field were directed by Clausen. Substantial contributions to the development of the final questionnaire and its analysis were made by Leta Adler, William McPeak, Shirley Star, and Edward Suchman.

would materialize with heavy demobilization. This was particularly true of agencies whose programs tied in with legislation like the Servicemen's Readjustment Act of 1944. The Department of Agriculture, for example, was concerned with providing information on farming to military personnel interested in buying a farm, to insure that such plans might have their basis in a realistic appraisal of opportunities. The Department of Commerce and the Smaller War Plant Corporation were particularly interested in the plans of the men and women in service who might try to start business enterprises. The United States Office of Education sought information on the plans of military personnel for school or college attendance after their discharge. Selective Service had a need for information on the extent to which selectees and volunteers alike intended to return to their previous employers. The Veterans Administration, charged with administering most legislation affecting veterans, including the educational and loan provisions of the GI Bill of Rights, would require estimates of soldiers' intentions with respect to these provisions. The War Department itself was concerned with getting some indication of the intentions of its personnel to remain on active duty after the war.

Here, then, was a problem in prediction or, more accurately, here was a series of problems in prediction. Primarily, interest was focused on answering the question, "How many?" Alternative approaches were possible. One might attempt to estimate the proportion of all men in the Army who would follow this or that possible course of action, basing such an estimate on a survey of the plans and characteristics of a cross section of the Army population. Or, one might approach the problem from the other side and estimate the relative opportunities which might be available to the veterans after the war and the probable pressures upon those opportunities, attempting to make some quantitative judgments.

Aside from the practical problem of arriving at an estimate of "how many," there were a number of rather intriguing theoretical problems which emerged as the studies progressed. There was the problem of assessing statements of intention applied to a vague future. Could a soldier's statement that he expected to go to school after being discharged from the Army be taken as evidence that he was really likely to go? More generally, under what conditions are statements of intention predictive of future behavior?

Most predictive studies in the social sciences have been based largely on the deterministic effects of "background characteristics."

How important would such characteristics prove for estimating *who*, as well as *how many*, would go to school? Where background and intentions clashed as predictors, which would be the most potent?

What kinds of questions and of interview or questionnaire design would give the most trustworthy results? Could any reliable estimates be made prior to demobilization? How much effect would a change in situation—from "soldier without immediate prospect of discharge" to "separatee 24 hours away from veteran status"—have on postwar plans? How would a change in the situation encountered after discharge affect the actual carrying out of plans?

As is almost always the case in social science research where a large number of uncontrolled variables are involved, the studies reported here did not yield unequivocal answers to the questions posed above, but they did yield tentative answers to some of them and suggested hypotheses for future testing in the case of others. The predictions based upon these studies were not uniformly accurate. Soldiers' plans, as analyzed and evaluated, served as a basis for reasonably good predictions of the number of men who attended full-time school, returned to farming after the war or returned to their previous employer; but plans for business ownership and government jobs proved to be less closely related to actual performance.

The following sections of the present chapter describe the general problems faced in carrying through the research program centered on soldiers' postwar plans and take up in detail the procedures used and the results achieved in attempting to estimate how many veterans would seek to attend school or college after their military service. In Chapter 16 we shall first examine two other instances of relatively successful prediction—extent of return to previous employer and extent of return to farming—and then consider two less successful attempts—predictions of business ownership and of governmental employment—seeking to analyze the reasons for the relative accuracy or lack of accuracy in these predictions.

The Problem of Ascertaining Soldiers' Plans

Let us first consider the problem of trying to ascertain the plans soldiers had for their civilian careers at the time of the pretests for the first major survey on this subject in the spring of 1944. Four fifths of the men in the Army had been in service for more than a year; roughly a third had been in service for more than two years.

But the cross-channel invasion was yet to be launched, and the prospects for an early return to civilian life by any substantial number of soldiers were very dim considering the additional fact that Japan still controlled the Philippines and most of the other territory she had wrested from the Allies in the first year after Pearl Harbor. Given the effect of two years of Army life—two years out of contact with civilian jobs or school—and the uncertainty of the future, how many men would have plans clearly enough in mind to express them in response to a survey? Would not many soldiers simply put down the first thing that came into their heads, or equally bad, would they not rise to the bait of any suggestions contained in questions as to their intentions? There was still another possibility—that in the time they had for dreaming, during the dull routines of Army life, they would conjure up wonderful images of a postwar career which they could not in reality hope to achieve. These dreams, then, and not reasoned plans, would be caught by the survey.

But even if accurate indications of their plans could be drawn from the soldiers, of what value were the intentions of men in such a state of uncertainty, so subject to new and changing conditions before they could hope to carry out their plans? The Servicemen's Readjustment Act of 1944, popularly called the GI Bill of Rights, would be passed in June, but not more than a tiny fraction of the men knew the provisions of this bill at the time of the first survey. Reemployment, for those who had prior to induction held nontemporary jobs, was guaranteed by the Selective Service Act, provided that the job still existed. But no one in Washington could say how many servicemen would qualify for reemployment under the provisions of that act. Some economists were predicting eight million unemployed during the demobilization period. Few men in service knew what they could expect to find after discharge.

The Background of Soldiers' Plans

At least four fifths of the men who entered the Army after September 1940 came more or less directly from civilian jobs which they had occupied as employees. Roughly a tenth entered the Army within a few months after completing or leaving school. A smaller proportion had been self-employed in small businesses or in agriculture, leaving perhaps 2 or 3 per cent who had experienced a considerable period of unemployment prior to their enlistment or induction. Among the men who had been employed, many had

held jobs for only a year or two, or even less. A considerable proportion of the men inducted during the latter stages of the war had been employed in war industry.

For some men, Army service represented a disruption of a planned career—a disruption of schooling, of apprenticeship, of growth in a chosen field of endeavor. For other men, service meant a break in a boring routine, a chance to get out of a job into which they had perhaps drifted by taking the course of least resistance. The return to civilian life might mean either taking up where one left off or making a fresh start. For many soldiers, perhaps for a majority, the break caused by Army service meant a chance to evaluate where they had gotten and to reconsider where they were going. Most men had worked prior to entering the service and might be expected to work after their discharge, but beyond that, the picture was not clear. Some men felt sure they would go back to their old jobs; some felt sure they would not; many were between these extremes.

One of the first facts that forced itself into view in the early pretesting of questions relating to soldiers' postwar plans was that any man could be classified as to his interests and plans along a number of dimensions. Picture a series of scales each representing interest in a specific type of postdemobilization activity under consideration—for example, interest in going to school, in going back to work for one's preservice employer, in starting a business, etc. A soldier planning to return to work for his preservice employer might rank high on a scale of interest in this dimension yet still be considering going to school if government aid were available to him, and hence rank far from the bottom with respect to interest in full-time school. Most men, in fact, even though fairly sure of their plans, were at least considering various alternatives depending on the materialization of a number of contingencies.

This was not the only problem of differing dimensions, however. Most men planning to work as employees after discharge were oriented toward a particular employer or industry, toward a particular occupation, or even toward a particular occupation in a particular industry. But some men were less concerned with occupation and industry than with the location of their employment, the salary at which they would work, the independence they would have—or security, or the compatibility of the work, or the opportunity for advancement. While one might safely categorize a large group of soldiers as planning to work for an employer, it was not possible to

make any adequate classification of the details of their plans because they could not be evaluated in terms of any single dimension. Adequate classification could be made only when a dimension was clearly specified and the interests of all soldiers in the survey sample were evaluated with respect to that dimension. Further, if the dimension was one arbitrarily imposed without regard to the soldier's frame of reference, there would be less likelihood that the classification would be meaningful or predictive of subsequent behavior.

There were, then, a number of serious problems to be overcome merely in order to arrive at an adequate classification of the plans which soldiers had. Beyond this, there was the crucial question of whether or not useful predictions could be based upon a knowledge of soldiers' plans.

SECTION I

THE PREDICTION PROBLEM

The attempt to predict what veterans would do after discharge, from a knowledge of their plans expressed prior to discharge, posed a type of prediction problem which has not received much attention in the literature of social science methodology.

To the writer's knowledge, there exists no adequate classification of the types of prediction which have been or might be undertaken in the social sciences. It is possible to classify prediction studies along a number of dimensions relating to methodology, degree of specificity, type of activity predicted, etc. One may, for example, distinguish between "absolute" predictions (i.e., involving unequivocal specification of performance for the individual case or the course of group action) and actuarial prediction (involving specification of the probability that individuals of stated characteristics will behave in a given way). Again, one may distinguish between prediction based at least in part upon individual intentions and the individual's conception of his role—self-prediction—and prediction based upon the judgments of others or upon factors in the background of the individual; between predictions based upon statistical analysis of the past relationship of various factors to the criterion which is to be predicted, and predictions based upon the construction of ideal types or coherent behavior systems not involving projection from past experiences under similar circumstances.

For purposes of fitting the present study into a larger frame of

reference, it may be useful to classify prediction studies according to the specificity and type of activity predicted, as follows:

A. *Predictions for the individual case*

1. Performance—level of attainment, adequacy of adjustment, etc.—in a specified role or task (here would be included the bulk of studies on prediction of marital adjustment, academic or occupational success, parole behavior, etc.)
2. Prediction of decisions, choices, preferences in selection, etc., usually involving a knowledge of the individual and of the situation in which the decision or choice is to be made. Such studies might be subclassified according to whether they relate to:
   a. Unique situations and one-time decisions (e.g., occupational choice at time of leaving school or at time of demobilization from the Armed Forces)
   b. Extended courses of action or developmental sequences (e.g., changing interests and activities with aging)

B. *Predictions for groups of individuals*

1. Frequency of occurrence of specified behavioral characteristics in given groups (e.g., prediction of the number of juvenile delinquents likely to come from a given population segment)
2. Characteristics associated with choice of or success in a particular activity (e.g., prediction of the characteristics of personnel who might be secured for the Armed Forces by voluntary enlistment)
3. Reactions of members of a group to a particular situation or program
   a. Where focus is upon proportions reacting in a given way (e.g., prediction of attitude changes resulting from exposure to a program of propaganda)
   b. Where focus is upon the course of mass behavior (e.g., probability of race riots in connection with a proposed change in the pattern of relations between races)
4. Patterns of cultural change or end products of social processes (e.g., prediction of fashions, of the course of inventions, of the impact of technology on other parts of culture, etc.)
5. Trends in vital indexes, demographic characteristics, etc.

In so far as individual prediction is based upon an actuarial approach, the predictive process would involve type B–1 before predictions under A–1 or A–2 could be made. The methods used in "screening" personnel, accepting or rejecting individuals above or below a certain critical score, based upon a knowledge that groups of individuals so categorized have a high expectancy of success or failure, may be considered one instance of this type. Where nonactuarial approaches are used, the predictive process for the individual case will usually differ significantly from that used in predicting for groups.

The great bulk of methodological studies in prediction for the

individual have dealt with problems where the objective was prediction of performance level or adjustment of the individual (A–1) through analysis of the past or prevailing frequency of high or low adjustment in various subclasses of the population studied (B–1).[2] The possibility of predicting directly for the individual, as a unique personality in a unique socio-psychological field, using intensive case history materials, has been stressed by Allport[3] and others, but relatively few examples of this type of prediction—*as prediction*—are to be found in the literature.

In the studies under consideration, several types of prediction were attempted. As was noted in the introductory paragraphs, the initial objective was to estimate *how many* veterans might be expected to make certain choices or follow certain courses at the time of leaving the Armed Forces. Thus, the primary classification would be type B–1. A second objective was to estimate *what kind of veterans* would enter specified activities. What proportion of veterans returning to school would be married? What proportion of veterans who wished to farm would have experience in farming? What kinds of veterans would try to start businesses of their own, and what sort of financial resources would they bring to the enterprises? These were important questions to answer in order to provide realistic information and other assistance to veterans. These would be predictions of type B–2.

Most predictive studies in the social sciences have involved generalizing from past observations of the relationship between a set of predictive items and the criterion to be predicted. But in the present instance no previous experience existed to indicate which items were truly predictive. Therefore, the approach had to be in terms of an analysis of the dispositions and intentions of individuals as affected by military service, or based upon the projection of trends in activities pursued before the war by the wartime population of the Armed Forces, or upon judgments as to the opportunities which might be available to veterans after the war.

In so far as interest was confined to forecasting how many soldiers would go to school, farm, or start businesses, the projection of past trends was feasible but not very promising. One might estimate school attendance, for example, by applying rates of school attend-

[2] The monograph by Horst et al., *The Prediction of Personal Adjustment* (Social Science Research Council, New York, 1941), dealt almost exclusively with this type of prediction.

[3] Allport, Gordon, *The Use of Personal Documents in Psychological Science* (Social Science Research Council, New York, 1942), pp. 154–161.

ance by age and level of previous education for the civilian population to an assumed age by education distribution of veterans at time of discharge. Or one might utilize census data on self-employment by age groups to estimate the probable increase of self-employment among the Army population had there been no war. But such methods must of necessity ignore the fact that the war had in all probability modified previous trends and relationships. With government aid promised to the veteran who wished to return to school, it seemed idle to assume that the number who would return to school would merely reflect past trends and would therefore be markedly less than the number who left school to enter the service.

For this reason, the approach to prediction for the whole veteran population was made through an attempt to predict for the individual. If one could make a sufficiently intensive analysis of the plans and characteristics of individual soldiers, so as to have confidence that one could make certain qualified predictions for the individual, it might be possible to build up estimates for the service as a whole. The problem was, therefore, viewed originally as one of making the best possible assessment of soldiers' plans, evaluating those plans in the light of their definiteness and realism and in terms of the characteristics of the planner, and then of presenting these data in sufficient detail and with necessary qualifications so that the consumer could consider them along with other data available in his specific field of interest.

In presenting the data on soldiers' plans, the policy was followed of projecting the sample estimates to the total population of male personnel in the Army as of the date of the survey. No assumptions were made about men already discharged or men yet to be inducted into the service. It was recognized that most people interested in such data would like to have estimates for the total veteran population, but such a population could not be clearly defined until the war was over.

At the time the first Army-wide estimates of soldiers' plans were prepared in the summer and fall of 1944, over a million men had been returned from the Army to civilian life, predominantly by reason of physical and mental disabilities. Many more men would be discharged before general demobilization would begin. In large proportion they would be eligible for the same benefits as veterans later to be discharged. But one could not safely assume that men previously discharged had the same orientation toward postwar jobs and school as did the group surveyed, and one could not assume

that men to be discharged by reason of disability would sustain the plans they had earlier expressed. Therefore, one could generalize survey findings on soldiers' plans only to the population actually represented by the sample surveyed.

Further, it was stressed that data on plans were only one element to be considered in any prediction of what veterans would do after the war. If plans were not modified by changing circumstances prior to discharge they should at least indicate what veterans would *try* to do, and not necessarily what they would actually do.

Nevertheless, upon their release, the data on soldiers' plans were frequently quoted and requoted as predictions of the number of veterans who would attend school, start small businesses, and return to farming, without reference to the fact that the survey findings represented only a portion of the total population of veterans (i.e., only men in Army service at the time of the survey) and without reference to the effect of the economic conditions to be encountered either on the sustaining of soldiers' plans or on the success soldiers would have in carrying out those plans after discharge.

If some consumers used the data indiscriminately, others who had a need for estimates of what veterans would do dismissed data on soldiers' plans as irrelevant. "Soldiers don't know what they want to do," argued one official, "I've talked with dozens of them and most of them haven't the foggiest notion of what they will do after they're discharged."

But despite indiscriminate use of the data in some quarters and skepticism as to their value in other quarters, the release of a series of reports based on the first Army-wide survey of soldiers' plans was followed by many requests for further details and for continuing studies to indicate possible shifts in soldiers' intentions. The planning of subsequent studies in this field recognized much more fully than had the initial planning that the prediction problem would have to be faced squarely in future work.

As a coordinate part of a world-wide survey of soldiers' plans scheduled for the summer of 1945, it was decided to survey the plans of men being discharged to civilian life and then to follow up these men several months after discharge in order to learn whether or not they had attempted to carry out their plans. Once the relationship had been established between plans or intentions expressed while in service and actual postseparation performance, it was hoped that more accurate forecasts could be made from data on the plans and intentions of the total population.

Although the agencies which requested data on soldiers' plans did not always know just what kind of data they wanted or how they might use the data, it developed in the course of contacts with them that by and large they wanted ready-made predictions and not merely a set of facts about soldiers' plans. In general, they wanted predictions of the amount of pressure on their programs which would come relatively early in demobilization or at least relatively soon after heavy demobilization had taken place. How rapidly should facilities of the Veterans Administration be expanded to meet the demands for educational benefits? How much money should be budgeted for the immediate demobilization period to cover anticipated needs? What kinds of information about job opportunities and problems of self-employment should be made available to soldiers in the demobilization process in order to facilitate the smoothest possible transition to civilian life?

To provide assistance in answering such questions, it seemed desirable to aim at predictions of the number of soldiers who might reasonably be expected to try to enter the activities under consideration. That is, one would not attempt to predict the number of soldiers who would actually enroll in colleges or succeed in starting businesses, but rather the number who might be expected to make an effort to enroll in college or to get a business into operation. Actual enrollment in college would depend in part on the availability of space in colleges, on the availability of housing for married veterans in college towns, etc. Success in starting a small business might depend on the availability of materials or of a site and a building. For planning purposes, interest was therefore primarily focused on *demand* for facilities.

In order for data on soldiers' plans to serve as a basis for accurate predictions, the following conditions would have to obtain:

1. It should be possible to classify a large majority of all soldiers into definite categories with respect to their plans. In so far as men could not be specifically classified—the men with vague plans or alternative plans or no plans at all—it should be possible to specify the general employment status which they would be most likely to choose and the probability that the undecided group would shift in large proportion to any other specific activity under consideration. It is apparent that if 8 per cent of the men planned to attend school but 20 per cent were completely undecided in their plans, an estimate of the number likely to attend school could be based upon

plans to attend only if one could feel confident that few of the undecided group would ultimately decide to go to school.

2. The classification of plans should represent accurately the more or less organized attitudes and intentions—the orientation—of the individual classified, indicating a tendency to act in a given (general) way under a given set of stated or assumed conditions. In other words, as of the time of initial study the classification of plans should possess both stability or reliability and internal consistency (given a definition of plans in terms of the organization of attitudes toward a given goal).

3. There should be no marked change in personal characteristics or social conditions which would influence the individual substantially to modify his plans between the time of the survey of plans and the time of return to civilian life. In other words, there should be no materialization of unexpected contingencies which would lead to a change of plans.

4. Apart from changing conditions, plans should be based upon a sufficiently realistic evaluation of the situation and problems to be encountered after discharge so that the confronting of that situation would not lead to the immediate abandonment of plans.

It was apparent from the start that not one of these conditions would be *completely* realized. It was known that some men with vague plans or no plans at all must be considered potential aspirants in almost every field for which estimates were desired—although an analysis of the characteristics of the undecided group suggested that a large majority would seek some sort of semiskilled or unskilled job with an employer.

It was known that some of the men classified as having definite plans to enter this or that field were really just daydreaming and that others were occasionally misclassified. But, the intensive pretesting had yielded a fair degree of confidence that our classifications were valid classifications of what men actually intended to attempt and expected to do.

It was known that some men were making unrealistic assessments of conditions they would encounter or of aids which would be available to them—but again it appeared that *most* men made fairly sober assessments.

Finally, it seemed obvious that some men would modify their plans as a result of subsequent experiences while in service, and also that the conditions of the job market and the terms of aid available

to veterans would almost certainly be in a state of flux between the time of our surveys and demobilization.

It is readily apparent that validation of the data on plans and intentions and successful prediction from those data are two quite distinct problems. If intentions expressed while in the service may be regarded as representing an organization of attitudes toward a particular goal, it does not follow that any specific behavior pattern is a requirement for validation of the measure of intentions.

Thus, in relating performance data to plans data, we were not testing the validity of questionnaire responses as measures of plans and intentions but were testing a more general hypothesis—or rather a series of hypotheses—on the relationship of the questionnaire responses given under one set of circumstances to behavior under a different set of circumstances at a later date. It was believed that in so far as the conditions earlier specified did not actually obtain, some men with definite plans would modify those plans or for other reasons fail to attempt to carry them out, and other men with only vague plans might well decide to enter any specific field under consideration. The problem was to estimate in advance of any testing of the relationship just what proportion of the soldiers in any category would attempt to carry out their plans—and, hence, what cutting points should be used for prediction.

The effect of various contingencies on the extent to which plans would permit prediction of postseparation behavior would clearly be expected to differ for different types of activities. For example, an increase in the amount of aid available to veterans for education would presumably have a greater effect on the prediction of return to school than on prediction of any other activity, though all activities from which some soldiers switched to school would thereby be affected to some degree.

From interviewing and the use of contingency questions it was possible to assess crudely the relative importance that soldiers themselves attached to the materialization of certain contingencies, but the differential nature of the effects was such that there was no way of preparing overall estimates of soldiers' plans under stated combinations of contingencies.

The predictions that were made, then, from the plans expressed in answer to questionnaires, were in large part based on assumptions and hunches derived from intimate acquaintance with soldier thinking on the subject of postwar plans and from a much sketchier

knowledge of conditions and problems likely to be encountered in the demobilization period.

These hypotheses, speculations, or assumptions were to be checked as regards the plans expressed at the time of discharge for men discharged in the early phase of demobilization. It was realized that any relationship established between plans and performance of this group would not, in all probability, remain invariant for later separatees, but this was to be the starting point for evaluating plans as a basis for prediction. Again, it was recognized that even if separatees' plans were predictive of their performance, the same might not be true of plans of men faced with the prospect of a year or more of additional service. In order to have any confidence in predictions based on plans of such soldiers, further evidence would be required that plans expressed in the absence of any expectation of discharge would be sustained up to and through the separation experience.

It was not feasible to follow up after discharge a group of soldiers who had been surveyed as to their postwar plans before the men had any definite expectation of discharge. But data on the relative stability of plans were secured, to some degree, by comparing plans of separatees with those of soldiers remaining on active duty, holding constant certain crucial characteristics of the two groups. Thereupon the separatees were followed up, after two to four months of civilian life, to find out how many attempted to carry out the plans they had described at separation. The technique used for the follow-up is briefly described at the close of the following section which deals with the general procedures followed in studying soldiers' plans.

SECTION II

APPROACHES TO ASCERTAINING SOLDIERS' PLANS

The interview. Interviewing soldiers with reference to the problem area under consideration before attempting to formulate a specific set of questionnaire items was a procedure regularly followed in major surveys, especially in studies which tapped areas not already explored in detail by previous surveys. In those instances where the objective of a study was to ascertain attitudes toward a particular problem, situation, or group, the interview might be fairly narrow in scope and the relevant data might be obtained by

a small number of questions. In questioning soldiers as to their postwar plans, it was found that a few general questions would give a fairly good indication of the general trend of the individual's thinking—whether he planned to return to his previous job, seek a new job, start a business enterprise, or attend school—but that a check on the degree to which plans were thought out and were realistic for the individual planner required a good deal of probing into the details of those plans. This finding suggested that the interview was perhaps the only suitable method for arriving at an accurate classification of soldiers' plans. In an interview it was possible to "feel out" the soldier, prodding to find out how well crystallized and how realistic his plans seemed, suggesting possible problems which might prevent the achievement of planned goals and getting the reactions of the respondent to those problems. Individual A, the soldier who stated categorically that he would have a restaurant of his own after the war, might hesitate and backtrack when asked about the cost of equipping a kitchen or the problem of locating a suitable site and getting into operation with a minimum of capital while supporting a family. These questions had to be asked in order to evaluate the plans of Individual A, but they would be completely irrelevant to Individual B who planned to return to school. And in an interview, only the relevant questions need be asked, i.e., those questions relevant to filling in some of the details of the specific plan or plans of the respondent and checking his reactions to alternative courses which he might reasonably be expected to follow.

This gave the interview a tremendous advantage over the questionnaire, which could not hope to present the relevant questions for each individual soldier and could not prod and probe in the same sense as an interviewer could. But interviews are relatively costly, since they require a great expenditure of time by trained personnel, and any adequate classification and analysis of soldiers' plans would require a carefully selected sample of many thousands of cases. The need for a large sample is readily apparent if one considers the task of classifying and analyzing plans to enter any single area of interest to the study. Only a small proportion of the men in service had definite intentions to start a business, for example, yet it was desired not merely to estimate the magnitude of this proportion but also to establish what kind of men had such plans, what specific businesses they planned to enter, what experience they had had in such businesses, and what arrangements they ex-

pected to make in order to finance these enterprises. This could be accomplished only by taking a total sample sufficiently large to yield several hundred cases in each of the major areas of interest.

The small field staff of the Research Branch, relative to the volume of research demands, precluded any possibility of doing an adequate study of soldiers' plans by any method other than group administration of questionnaires. Thus the problem became one of devising a questionnaire which would both give an adequate classification of soldiers' plans and intentions, by broad areas, and provide sufficient detail to evaluate the plans and make the data useful for planning purposes.

The questionnaire. Like most questionnaires, the final form used for the study of soldiers' plans was not arrived at by a single effort but was the result of considerable testing and revision. In fact, in the amount of testing and reformulation, this questionnaire was perhaps exceptional. A brief description of the steps in the development of the final form of the questionnaire will indicate the reasons for this.

The development of the questionnaire through pretesting. The first questionnaire pretest in the general area of postwar plans was primarily an attempt to secure a set of questions which would assess the amount and conditions of interest in remaining in the Army after the end of the war. It also tested a selection of items from a previously developed battery designed to rank soldiers according to their interest in postwar education, and contained several questions relating to general job plans.

A battery of questions which permitted a rough ranking of respondents on interest in postwar Army service was developed, and the items relating to postwar education again provided a reasonably good ranking with respect to this area. It was found, however, that a number of individuals who were completely consistent in saying that they would attend school, in answer to a series of questions relating only to school (or were likewise consistent in expressing plans for an Army career), said they would do something else when presented later in the questionnaire with a general question permitting the selection of one of a series of alternatives. These alternatives included plans to do the same kind of work as the soldier had previously done, plans to do a different kind of work, plans for school, and plans to remain in the Army.

This seemed to indicate either a lack of any stable plans or a tendency of some men merely to respond to suggestion when presented

with a direct question or series of questions relating to plans with respect to a single course of action. It appeared also that question order was extremely important in determining responses to single items or to whole scale batteries. What would happen if the question involving alternatives were to precede the questions relating to school or to an Army career, or if question batteries were broken up and distributed throughout the questionnaire?

A second pretest was designed to test the effects of context on response, using four questionnaire forms which were administered interleaved to approximately two thousand men. Results were not conclusive, in the sense of showing uniform shifts when direct items or scales were preceded or followed by general questions presenting or implicitly containing alternatives, but a number of marked shifts were observed. In general, however, a direct question relating to plans for a single course of action was found to draw more affirmative responses for that course of action than a question which contained that course as one alternative, regardless of order of presentation.

The second pretest also contained a number of questions relating to interest in farming, in self-employment, and in migration. Relatively high consistency was found in the expression of plans for farming, which came predominantly from men with a farm background. Surprisingly high aspirations for owning a business were found, though in large proportion the aspirants indicated other plans when presented with alternatives. Expression of plans for migration seemed likewise not wholly consistent either within the set of items relating to migration or in comparison with occupational plans.

A final pretest was set up to sharpen the items used and to achieve an item order which would minimize inconsistency and effects of suggestion. Questions on plans were preceded by a set of questions on preinduction employment experience and were introduced by a write-in question relating to general job plans. This form was on the whole found satisfactory and, with some modifications of individual items, was administered to more than twenty-three thousand officers and enlisted men in the continental United States and overseas in the summer of 1944.

The final questionnaire used in the summer of 1944 was designed not merely to classify individuals according to their relative interest in following a particular course of action after the termination of hostilities, but to analyze the details of their plans. It was aimed

at securing reasoned evaluation of plans—recognizing the limitations of the questionnaire for achieving this purpose—rather than reflections of aspirations. It sought to minimize the effects of suggestion by first getting the respondent to write out his plans in answer to a general question which contained no suggestive elements.

The adequacy of a questionnaire to fulfill the requirements for which it is designed is not merely a matter of asking the right questions. An integral part of the problem of survey design is the plan for analysis of the data secured. In the case of the study of postwar plans it was found that differing approaches to analysis could yield markedly different estimates from the data. The analysis plan for the first major study of postwar plans was in many ways awkward. The codes set up for classifying data contained in the write-in questions were exceedingly complicated because of the number of dimensions involved. Therefore, even before the results of the July 1944 survey had been fully reported, additional pretesting was begun to try to simplify the analysis plan and to secure additional detail on plans for business ownership and school attendance. Further experiments were conducted on the effects of context, and an attempt was made to arrive at a sequence of items which would minimize suggestion without requiring use of a general write-in question. It was found, however, that the use of the general write-in question as a lead-off item and the general sequence of the 1944 version gave results which seemed less affected by suggestion than did any of the alternatives tried. The final questionnaire used in the summer of 1945 was, therefore, basically very similar to the form used a year earlier, though expanded in some respects (plans to work for an employer, plans to own a business, plans for school) and contracted in others (plans for further Army service, plans to migrate). Perhaps the most important change was the addition of a question toward the close of the questionnaire asking for a reconsideration and final evaluation of the general alternatives of employment status.

The preceding description of the steps in the development of the two forms of the questionnaire which resulted in estimates of soldiers' plans has merely touched upon some of the general problems encountered. It has been suggested that the adequacy of the questionnaire itself can be judged only in terms of the plan of analysis. The plan of analysis, however, was based in large part on assumptions with respect to the specific fields of interest covered by the questionnaire. For this reason, further analysis of the problem of

ascertaining soldiers' plans will be directed to several of these fields of interest—particularly to the problems of ascertaining plans for school attendance, business ownership, and returning to previous employer. In order to avoid undue repetition at a later point, the use of the data for prediction purposes and the accuracy of the predictions made will be discussed along with the technique of ascertaining plans for each area of interest. Before taking up these detailed analyses, however, a brief note on the follow-up technique is in order.

The problem of ascertaining performance. The general hypothesis to be tested by following up separatees was that men who expressed plans to follow a given course of action after their discharge—whether school attendance, returning to previous employer, self-employment, etc.—would actually attempt to follow that course of action. Therefore, the criterion to be evaluated was not the activity in which the veteran was involved at any particular time after discharge, but his efforts to enter a given field within a few months after leaving the service.

The method projected for the follow-up was to mail questionnaires to a sample of separatees from four widely scattered separation centers asking about their occupational activities, interests, and efforts. This questionnaire, again, had to be simple enough so that men with only a grade school education could answer it, yet detailed enough to permit evaluation of the criterion. A crucial problem, of course, was the definition of an "attempt" to carry out previously expressed plans. If a veteran actually succeeded in realizing his plan, or was actively engaged in pursuing it at the time of the follow-up, he could clearly be classified as having attempted to follow this particular course of action. If, however, he merely inquired about entering the field he had planned to enter, received a discouraging response to his inquiry, and gave up his former plans, it seemed doubtful that he should be classified as having made an attempt to realize the plans. For each of the fields with respect to which plans were evaluated, explicit criteria were established. These will be discussed in the sections relating to the several areas of interest.

The mail follow-up of men discharged in July 1945 yielded returns from 88 per cent of the group surveyed. In order to check on possible response bias, interviews were conducted with the bulk of the nonrespondents from two of the four separation centers. No response bias was apparent.

A second follow-up study was conducted among men separated

at the peak of demobilization (December 1945) in order to check on the effect of a markedly changed economic situation on the previously observed relationships between plans and performance. Again a response rate of 88 per cent was secured by mail questionnaire alone. Although a slight educational bias was observed, the questionnaires returned are believed to be sufficiently representative of the total group surveyed to provide acceptable evidence.[4]

SECTION III

POSTWAR EDUCATIONAL PLANS OF SOLDIERS

Somewhat less than a tenth of the Army's enlisted men and about a fifth of its officers were full-time students just before they entered the service. Some of these men had just completed high school or college when they enlisted or were inducted. Others cut short their education to go to war. Still others had left schools and colleges to take jobs in war industry. Cross-sectional surveys indicated that about 13 to 15 per cent of the total Army population at the time of peak strength had been out of school for less than one year when they entered service. Had there been no war, it is likely that the number of these men in school would have declined sharply with each year of aging, as some finished high school, some finished college, and some simply dropped out of school. It is virtually certain that not more than 5 per cent of the men demobilized from the Army during 1945–1946 would still have been in school at that time had there been no war.

On the other hand, it is probable that many of the Army's officers and enlisted men who had graduated from high school and then taken jobs would have liked to attend college if financial obstacles had not seemed too difficult to surmount. Given financial aid, they were potential students again. Other men, who had dropped out of school because they were not sold on the value of further education, saw in the Army the enormous advantages secured by their better educated fellow soldiers. Given another chance, they might choose to better their competitive position by seeking further schooling.

Because of the possible effects of government aid for schooling and the effects of Army service itself on an interest in further education, the plans of soldiers to return to full-time school were sought as an indication of the probable rate of return.

[4] The techniques used in the mail follow-ups have been described in an article by John A. Clausen and Robert N. Ford, "Controlling Bias in Mail Questionnaires," *Journal of the American Statistical Association*, Vol. 42 (December 1947), pp. 497–511.

Ascertaining Plans for School

The initial survey relating to plans for school was conducted nearly a year before the Servicemen's Readjustment Act of 1944 provided for aid to veterans wishing to return to school. The function of the survey was to provide an estimate of the proportion of soldiers who might be expected to return to school even without a program of federal aid, and an estimate of the proportion who might go if such aid were available. To this end a set of questions was developed relating to interest in returning to full-time school under various contingencies. These questions were carried toward the close of a questionnaire dealing primarily with problems of morale and adjustment. The sequence of questions on interest in postwar education followed immediately after a question on general occupational interest. The items with which we are chiefly concerned are given below, along with the percentage distribution of respondents to each category. (Question numbers are those in the original survey.)

86. If you could do what you wanted, would you like to go back to full-time school or college after the war? (Check one)

Yes	35%
No	48
Undecided	15
No answer	2
	100%

87. Regardless of what you *want* to do, do you think you will actually go back to full-time school or college after the war? (Check one)

Yes	8%
No	77
Undecided	12
No answer	3
	100%

88. Do you think you would go back to *full-time* school or college after the war if the Federal government or your state government were willing to help pay your way? (Check one)

I expect to go back to school whether or not the Federal or state government helps me	8%
I will go back if the Federal or state government helps me	23
I will not go back, even if the Federal or state government would help me	38
Undecided	26
No answer	5
	100%

90. If you were offered *a good job* at the end of the war, do you think you would turn it down and go back to *full-time* school or college, or do you think you would take the job? (Check one)

I would turn the job down and go back to school	6%
If I could get some help, I would turn the job down and go back to school	9
I would take the job and *not* go back to school	69
Undecided	11
No answer	5
	100%

91. If you could get *some kind of job* after the war, but *could not get a good job*, do you think you would turn it down and go back to *full-time* school or college, or do you think you would take the job? (Check one)

I would turn the job down and go back to school	20%
If I could get some help, I would turn the job down and go back to school	18
I would take the job and *not* go back to school	42
Undecided	14
No answer	6
	100%

92. If you couldn't get any job at all after the war, do you think you would go back to *full-time* school or college? (Check one)

I would go back to school	21%
If I could get some help, I would go back to school	27
I would *not* go back to school	30
Undecided	17
No answer	5
	100%

93. If the war lasts longer than you expect, will that make any difference in whether you go back to *full-time* school or college after the war? (Check one)

I don't think I will go back to full-time school or college	40%
I will go back to full-time school or college, even if the war lasts longer than I expect	10
I will probably *not* go back to full-time school or college	27
Undecided	17
No answer	6
	100%

94. If you could do what you wanted, which would you rather do after the war—go to full-time school or college, or go to part-time school or college? (Check one)

I would not like to go to school or college at all after the war	35%
I would like to go to full-time school	18
I would like to go to part-time school	29
Undecided	12
No answer	6
	100%

95. Regardless of what you want to do, do you think you will actually go to full-time school or college, or go to part-time school or college after the war? (Check one)

I don't think I will go to school or college at all after the war	52%
I think I will go to full-time school	8
I think I will go to part-time school	19
Undecided	16
No answer	5
	100%

A careful scrutiny of the "marginals" reported for each question indicates considerable variation of the proportion interested in full-time school under varying contingency specifications. While 35 per cent of the soldiers said they would *like* to go to full-time school or college after the war in answer to the first question in the sequence, only 8 per cent said they thought they would actually go in answer to a direct question which mentioned no contingencies. Under the most favorable contingency from the point of view of school attendance—government aid available for education but no jobs available—48 per cent of the men said they would go to full-time school (Q 92). Thirty-eight per cent said that with aid available they would go to full-time school, if they were able to get some kind of job but not a good job, while 15 per cent said they would turn down a good job to go back with government aid. In considering each of these job contingencies, we find that roughly half of those who said they would go to full-time school indicated that they would do so only if government aid were available. Mention of the single contingency of government aid, with no mention of the job market (Q 88) led 31 per cent of the sample to say they would return if such aid were available.

The proportions of men who said they would go back without aid, and the proportions who said they would go back if aid were available are summarized for various contingencies in Table 1.

When the job and government aid contingencies were dropped, 10 per cent of the men reported that they would go back "even if the war lasts longer than I expect" (Q 93), and without mention of contingencies, 8 per cent again said they would actually go back to full-time school (Q 95). A consideration of these two figures gives a clue to the interpretation of responses to contingency questions. Interviews with enlisted men had established the fact that many men who hoped to return to full-time school were worried that they would be too old to return if the war were to last very

long. Yet responses to the unfavorable contingency actually showed a higher proportion reporting plans for full-time school than when no contingency was presented. However, the unfavorable war-duration contingency followed the most favorable of all situations presented—government aid for education and no jobs available—and it appears that in answering this question on the effect of a long war, some men carried in their minds the previously presented situation. That is, they were considering all three factors—aid, unfavorable economic situation, and duration of war—in answering the question which was intended merely to present

TABLE 1

PLANS FOR FULL-TIME SCHOOL EXPRESSED IN ANSWER TO QUESTIONS PRESENTING VARIOUS CONTINGENCIES RELATING TO THE AVAILABILITY OF JOBS AND OF GOVERNMENT AID

Question number	*Job contingency*	*Say they would go without aid*	*Say they would go with aid**
88	None mentioned	8%	31%
90	Good job available	6	15
91	Some kind of job available	20	38
92	No job available	21	48

* Includes those who would go without aid

the third of these. By the time they got to the final question, asking for a self-prediction without specification of contingencies, it would appear that they were making a more general evaluation of the situation.

It must be realized that whether or not contingencies were mentioned in the questions asked, the plans expressed by the soldiers were to a considerable degree contingent plans. There is evidence that, by and large, responses to contingency questions do not represent careful consideration of the specific contingencies presented, but merely reflect general reaction to a more or less favorable situation than that assumed (in a given context) if no conditions are mentioned.

The data previously presented on plans for school attendance, with government aid and without such aid under various job contingencies (Table 1), show marked shifts in the differential effects of aid under various circumstances. It is apparent that many men must have found it difficult to evaluate the various contingencies in a consistent fashion. This is not surprising, for "a good job"

and "some kind of job [that is] not a good job" are concepts not easily evaluated, and the same is true of the availability of an unspecified amount of government aid to weigh against job contingencies. It appeared, then, that the condition of the job market and the availability of government aid might have great influence on the extent to which veterans would return to school, but it seemed doubtful that soldiers' reports of their plans under various contingencies could be accepted at their face value.

In addition to the problem of interpreting the answers to contingency questions there were a number of inconsistencies which were

TABLE 2

INTERRELATIONSHIP OF RESPONSES TO TWO ALMOST IDENTICAL QUESTIONS ON EDUCATIONAL PLANS

(Question 87 and Question 95)
(Per cent)

95. Do you think you will actually go to full-time school *or* to part-time school?	87. Do you think you will actually go back to full-time school?			
RESPONSE CATEGORY CHECKED	RESPONSE CATEGORY CHECKED			
	Total	*Yes*	*No*	*Undecided or no answer*
Total	*100.0*	*8.2*	*76.8*	*15.0*
Full-time school	8.3	4.9	2.0	1.4
Part-time school	19.3	2.3	13.4	3.6
Not planning to go to school	51.4	0.4	48.3	2.7
Undecided or No answer	21.0	0.6	13.1	7.3

difficult to explain. For example, more than a third of the men who said they would turn down a good job and go to school even without government aid failed to predict that they would go when asked what they thought they would actually do. Perhaps the most startling evidence of inconsistency came from almost identical questions asking what the soldier thought he would actually do, separated in the questionnaire by seven items dealing with various contingencies. The cross tabulation of these two questions is given in Table 2.

It will be noted that the two questions give almost identical estimates of the number of soldiers planning to attend full-time school —8.3 per cent and 8.2 per cent—but that only 4.9 per cent are consistent in expressing such plans to both questions. It appears that some men who answered question 87 affirmatively were really inter-

ested in part-time school, but, given no opportunity to express interest in part-time schooling, they reported plans for full-time school in order to avoid seeming negative about any further education. But it is less apparent why 3.4 per cent of the men were negative or undecided in answering question 87, yet said they expected to go to full-time school in answer to question 95. Analysis of the relation between responses to those two questions and responses to the intervening questions reveals two possible explanations, however. First, there were a number of men with only grade school education who checked in answer to one or the other of these questions that they expected to attend full-time school although their response to the initial question in the sequence (Q 86) indicated that they had little interest in or desire for further schooling. These men with little schooling were apparently quite suggestible when asked a whole series of questions about school attendance. Second, there were a number of men who expressed themselves as undecided or said they did not expect to attend school in answer to question 87, but who said that they would go if government aid were available (Q 88). Having considered the possibility of such aid, these men may have been led thereupon to say they would actually go back, when given another chance to predict their future.

In view of these inconsistencies, how could the data best be utilized to give some estimate of the proportion who seemed likely to return to school after the war? Should responses to some one item be taken as the best indication of probable return to school, should complete consistency be required on all items, or could some other criterion be applied?

School Plans as a Scale Continuum

Responses to the battery of questions on educational plans were analyzed by the technique of scale analysis described elsewhere in this volume. The items were found to yield a fairly good approximation to a scale (according to the less rigorous criteria for scalogram analysis used in the early days of the Research Branch) after certain combinations of response categories had been made. From the data already presented on inconsistencies among responses, it will be obvious that the amount of scale error for responses to certain items was considerable.

An examination of the cross tabulation of questions 87 and 93 by educational groups (Table 3) is illuminating. To minimize inconsistencies resulting from errors in checking, the relationship be-

tween responses to the two questions is presented only for those men who had earlier checked that they "would like to go back to full-time school or college after the war." It will be noted that, relative to the proportion of each educational group who expect to return to full-time school, the contingency of a longer war than expected dissuaded a higher proportion of the men whose education stopped short of high school graduation than of those of higher educational status. More strikingly, however, among both of the lower educational groups the seemingly negative contingency actu-

TABLE 3

EFFECT OF LENGTH-OF-WAR CONTINGENCY ON EXPRESSED PLANS OF SOLDIERS WHO SAY THEY WOULD LIKE TO RETURN TO FULL-TIME SCHOOL, BY EDUCATIONAL LEVEL

(Per cent)

COMBINED RESPONSES TO QUESTIONS 87 AND 93	EDUCATIONAL LEVEL OF RESPONDENT		
	Less than high school graduation	*High school graduation*	*Some college*
Total who would like to return to full-time school	*100*	*100*	*100*
Q87: Think they will actually return	16	17	41
Q93: Say they will return even if war lasts longer than expected	(7)	(10)	(28)
Q93: Say they will not return if war lasts longer than expected	(9)	(7)	(13)
Q87: Think they will actually not return	84	83	59
Q93: Say they will return even if war lasts longer than expected	(10)	(11)	(3)
Q93: Say they will not return if war lasts longer than expected	(74)	(72)	(56)

ally drew a higher proportion of affirmative responses from men who had said they did not expect to return to school than it dissuaded from those who did plan to return. Among college men, on the other hand, the contingency was clearly evaluated as negative—it dissuaded 13 of the 41 per cent who had reported the expectation of returning to school, while it drew a positive response from only 3 of the 59 per cent who had not reported the expectation of returning.

One may hypothesize either that the less well-educated men were less consistent in their responses (perhaps because of suggestibility,

perhaps because of not having clearly formulated plans) or that the contingency had a different meaning to the college men than to those of less previous schooling. The soldier who needed only a year or two more of college to take his degree might be less discouraged from continuing on after discharge by having to spend a longer time in service than the soldier who had not yet begun college.

Even though the questions presenting various contingencies which might affect return to school after the war tapped more than a single dimension and thus did not yield a true scale, the technique of scale analysis did help to give a clearer picture of the patterns of relationship among response items than could be readily obtained by simple cross tabulation of questions. The combination of categories was facilitated by studying these patterns of interrelationship in the light of manifest content and the findings of intensive interviews.

By treating the questions and responses by the method of scale analysis it was possible to get a somewhat clearer picture of the patterns of response and the combinations of categories which seemed most reasonable in the light of their interrelationships. Moreover, by scoring individuals to the nearest scale type, it was possible to classify individuals who omitted one or two of the questions and to classify some of those who were inconsistent on only one or two items. That is, these individuals were classified as belonging to the scale type to which they could be scored with the least number of scale errors. Thus it was possible to rank individuals in a small number of groups, ranging from those who were consistently positive in expressing plans for school to those who had no interest in further full-time schooling, although rankings were crude and sometimes equivocal. But, given such a ranking of individuals based on their plans for school, how did one go about establishing a cutting point such that one would predict school attendance after discharge for men above the cutting point and predict nonattendance for men below the cutting point? Scale theory did not answer this question, and the cutting point actually used was based on considerations of consistency and manifest content of the items.

The Assessment of Plans for Predictive Purposes

It will be recalled that 8 per cent of the men surveyed in the summer of 1943 said they thought they actually would return to school. Nearly 7 of these 8 per cent were consistent in saying that they would like to go to full-time school, would go to school even without government aid, and would go back if offered any job but one they

considered a good job. Only about 4 of the 7 per cent, however, said they would return to school "even if the war lasts longer than I expect," and the same proportion (though not the same individuals) said that even if offered a good job they would turn it down and go to school without government aid. Only about 3 per cent said they would go under all contingencies presented.

It was felt that the 7 per cent who said they actually expected to return to school and were consistent in responding affirmatively to all of the questions except those dealing with the contingency of a good job and of a long war had rather coherent attitudes toward going back. This figure, therefore, was taken as an upper bound to the per cent who would go back if no aid were provided. But the analysis also considered the characteristics of the men who reported plans to attend school. It was felt that men who had been out of school more than a year before entering the Army, men who were married, and those over twenty-five years of age, would encounter great difficulty in attempting to go back to school, and therefore, even though their determination to attend might be great, relatively few could be expected to return. These men comprised nearly half of the 7 per cent who consistently expressed plans to return to school. Therefore, a lower bound of the estimate of school attendance by soldiers was set at 3.5 per cent, providing aid were not available to the soldier interested in returning to school. Because soldiers' thinking about the various contingencies of aid and the conditions of the labor market seemed rather vague, no estimate of probable enrollment under any particular set of contingencies was attempted. It was pointed out, however, that many of the men who said they might go to school if aid were available did not seem likely prospects for enrollment because they were over twenty-five, were married, or had been out of school for a number of years before they entered the service.

The first survey in the area of educational plans, then, indicated that responses to contingency questions could not be taken at face value, showed a large element of uncertainty and inconsistency in soldiers' plans for full-time education, and posed the problem of setting the cutting point for prediction when soldiers were ranked according to their degree of interest in full-time education.

Subsequent pretests revealed that a question which confronted the soldier with a choice of alternative courses of action—private employment, governmental employment, self-employment, school attendance, or continued Army service—gave a somewhat lower

estimate of the proportion with plans to attend school than did the cross tabulation of those questions relating solely to plans for school attendance which had previously been used as a basis for plans classification.

A pretest carried out among four-hundred white enlisted men at a single post found that of 44 individuals who had checked that they expected actually to attend full-time school in answer to the direct question on educational plans, only 24 also checked that they planned to attend school in answer to the question presenting alternatives. This suggested that reports of educational plans when the question context was solely educational interests were not comparable with reports when the context included alternative areas of interest. Men who were considering the possibility of attending full-time school after the war might say they actually expected to go to school when asked only about educational plans, but on considering the alternative of going to school or taking a job might decide that they would probably take the job.

Among men who checked that they planned to attend school in answer to the question involving alternatives, the vast majority also indicated that they would like to attend full-time school, that they planned to attend if government aid were available, and that they thought they actually would attend. Thus, this response could be used to obtain greater discrimination at the top of the crude scale ranking individuals according to their school plans. It served to categorize a group of soldiers who were not only highly interested in returning to school, but who were leaning more strongly in this direction than in any other direction with respect to their postwar plans and intentions.

When the final version of the survey was put into the field in June 1944, the following four questions were used to classify soldiers with respect to their plans for full-time school (marginals shown are based on responses of 19,556 white enlisted men surveyed in the United States and overseas, without correction for a minor sampling bias in level of education of respondents):

32. Do you expect to work for yourself or for someone else *right after you leave the Army?* (Check one)

I expect to work for myself	24%
I expect to work for the government (Federal, state, city, etc.)	9
I expect to work for a relative (father, uncle, etc.)	5

I expect to work for some other employer (company, person, etc.)	43
I plan to go to full-time school	9
I don't have any definite plans	7
No answer (including double checks)	3
	100%

42. Do you feel that you have had as much school or college as you want, or would you like to go back to school or college after the war? (Check one)

I have had as much school or college as I want	38%
I would like to go back to *full-time* school or college	19
I would like to go back to *part-time* school or college	39
No answer	4
	100%

43. It has been planned that the federal government will pay tuition and other educational expenses plus a living allowance of $50 per month to veterans who qualify and want to continue their schooling. Do you think you would go back to *full-time* school or college after the war, if this is done? (Check one)

I expect to go back to full-time school *whether or not* the government helps me	10%
I will go back to full-time school *if* the government provides this help	19
I will *not* go back to full-time school even if the government would help me	34
Undecided	30
No answer	7
	100%

44. Regardless of what you would like to do, do you think you will actually go back to school or college after the war? (Check one)

Yes, to *full-time* school or college	12%
Yes, to *part-time* school or college	22
No, I don't think I will go back to school or college	41
Undecided	21
No answer	4
	100%

These four questions again permitted a rough ranking, using scale analysis, giving the following ideal scale types:

1. Would like to go to full-time school
   Thinks he will actually go
   Expects to go whether or not the government aids him
   Chooses school among alternatives

2. Would like to go to full-time school
   Thinks he will actually go
   Expects to go whether or not the government aids him
   but
   Does not choose school among alternatives

3. Would like to go to full-time school
   Thinks he will actually go
   but
   Expects to go only if the government aids him
   Does not choose school among alternatives

4. Would like to go to full-time school
   but
   Does not think he will actually go full-time
   Expects to go only if the government aids him
   Does not choose school among alternatives

5. Indicates neither interest in going to full-time school nor plans to attend

Several additional scale types relating to interest in part-time school were also obtained, but need not concern this discussion. One nonscale type which was not large but was important was noted: nearly 2 per cent of the men deviated from type 1 in that they said they expected to attend full-time only if the government provided some aid. These men were considered as belonging with type 1 for predictive purposes, since government aid was assured.

Soldiers' Educational Plans in the Summer of 1944

Roughly 7 per cent of the survey sample of white enlisted men[5] fell into type 1 or the slightly deviant type as described above. These men were categorized as having definite plans for enrollment in full-time school. Another 4 per cent of the survey sample fell into types 2 and 3, which differed only in respect to the expectation of having government aid available. Since it was known that such aid would be available—the Serviceman's Readjustment Act of 1944 was passed as the survey went into the field—it was felt that types 2 and 3 should be combined. These men were thereafter categorized as having tentative plans to attend full-time school, even though they did not express plans to attend school when confronted with a choice of alternatives.

Four fifths of those with definite plans were high school graduates, under twenty-five years of age, and unmarried. All but 3 per cent

[5] Because white enlisted men comprised more than four fifths of the Army's total strength and because many of the pretests were conducted exclusively within this segment, the data to be presented will relate entirely to white enlisted men. Surveys were, however, conducted among officers and among Negro enlisted men.

had at least two of these characteristics. Roughly two thirds of the men with definite plans to return to school expected to attend college, while most of the remainder expressed interest in trade and business school courses. Less than 5 per cent were planning to return to academic high school work.

The findings with respect to soldiers' plans for postwar education were presented in detail in a separate report.[6] Since the interpretation of the data was crucial to their use for planning purposes, a full quotation of the interpretation offered in the final report is given below:

The method used in classifying men's educational plans is rigorous enough so that it may safely be said that very nearly all the eight per cent (of officers and enlisted men) classified as definitely planning full-time school really would go back to full-time school if the Army were demobilized today. Relatively few of the men with tentative plans for full-time school and the men who are now planning part-time school though they would like to go to full-time school can be expected to return to full-time school, especially in view of the fact that many of them have fairly definitely formulated employment plans. Tentatively, then, the best prediction of the number of men who will return to full-time school is in the neighborhood of eight per cent.

The Army is, however, being demobilized not today, but in the indefinite future. The question to which an answer is wanted is: How many men actually will return to school after demobilization? The figure of 8 per cent given above is the best answer which can be offered at the present time, but it is possible to point out factors in the situation which may alter this percentage.

Analysis of a study made among white enlisted men in the United States in October, 1944, indicates that their knowledge of the educational provisions of the "GI Bill of Rights" is still scanty and inexact. Less than a fifth of all the men know even three broad features of the provisions—that a man who was 24 years old when he entered the Army can get aid to return to school even though he had been out of school for several years before his induction; that the maximum amount provided for tuition and fees is $500; and that a single man is to receive a maintenance allowance of $50.00 a month—and only a third of the men definitely planning to return to full-time school know this much about the GI Bill. It is quite probable that men serving overseas have even less information about it.

Thus far, the GI Bill of Rights has not been an outstanding factor in the decision to return to school. In fact, only a quarter of the white enlisted men definitely planning to go to full-time school are depending on qualifying for benefits. (This, of course, does not mean that the remainder will not apply for benefits.) But among the men whose plans are not definite are many who could not go to school without aid and who are not yet fully informed about the existence of aid. Half the white enlisted men who are tentatively planning full-time school say that they need government aid in order to return, and all the men who would like to go full-time but are planning only part-time attendance think of the possibility of attending full-time school only in terms of receiving government aid. It is only reason-

[6] Originally issued as Report B–121 in October 1944 and reissued in March 1945 as Report B–133, "Post-War Educational Plans for Soldiers," one of a series entitled *Post-War Plans of the Soldier.*

able to suppose that some of these men will take advantage of the opportunity the GI Bill offers, once they know about it.

Economic conditions during and after the demobilization period will influence decisions about schooling, also. Widespread unemployment might lead men with no real interest in further education to return to school for the sake of government allowances, though the unemployment benefits provision should tend to minimize this. On the other hand, if good jobs at high pay are plentiful some men now thinking in terms of school may indefinitely postpone further education.

The date at which demobilization occurs may also affect men's decisions to return to school, since most men are implicitly making some assumption about the length of the war in deciding whether or not to return to school. The length of time one has been out of school, one's age and marital status, themselves interrelated variables, are highly related to going back to school If the war lasts beyond their expectations, many men now planning to go to school may have taken on the responsibility of wives and children, or may come to consider themselves too old or too long out of school to attempt to resume. Offsetting this to some extent will be the continuing induction of eighteen year olds, which tends to maintain the size of the younger age group from which most of the prospects for full-time school come.[7]

Later tabulations on the characteristics of the Army as of summer 1944 revealed that the sample used to prepare the basic estimates of soldiers' plans for school contained a slight surplus of men under twenty-five who were high school graduates or had started college before entering the service—men of the type most likely to plan to return to school. This bias led to an overstating of plans for full-time school. The original estimate for white enlisted men should have been closer to 6 than to 7 per cent. In subsequent studies, therefore, the estimates of school plans for the Army as a whole were derived by computing the proportions having such plans by age groups within the various educational levels. These rates were then applied to the known distribution of the Army by age and education.

Plans of Separatees

It was found that with increasing knowledge of the GI Bill of Rights, there was an increase in the proportion of soldiers planning to return to full-time school. In the summer of 1945, when surveys were conducted both among separatees and men remaining in service in the continental United States, it was estimated that about 8 per cent of all white enlisted men in service or in process of demobilization had definite plans to return to school. It was at this point that the quantitative evaluation of plans as a basis for predicting performance was undertaken.

[7] *Ibid.*, pp. 7–8.

The same basic questionnaire was used to survey plans of men remaining in service and plans of separatees. The questions on plans for school were those previously used, except for dropping the question on the effect of the availability of government aid, since this was no longer a contingency.

It will be recalled that the first step in evaluating plans expressed while in service, as a basis for predicting performance after discharge, was to ascertain whether or not such plans would be affected by the fact of separation. Since, however, soldiers remaining on active duty could not conveniently be followed up through separation, the test of sustaining their plans unchanged must be a comparison between plans of men remaining in service and plans of separatees, making allowance for certain differences in the characteristics of the two groups. It was assumed that if men of a given age and education expressed plans for school in the same proportion whether remaining in service or being discharged, then a follow-up of men surveyed at the time of discharge might be used to evaluate the predictive value of plans expressed by all soldiers surveyed.

Among men remaining in service in the continental United States in August 1945, 13 per cent expressed definite plans to return to full-time school or college. Among separatees surveyed in July 1945, at four widely scattered separation centers, 3 per cent expressed definite plans to return to full-time school or college. But the separatees were predominantly older men who had been out of school for some time before entering the service. The men remaining in service in the continental United States, on the other hand, included a large proportion of younger men who had been inducted directly from school, and many of these men planned to return. For men of comparable age and previous education, there was no significant difference between the school plans of separatees and those of men remaining in service.

By December 1945, approximately midway in demobilization, younger men without dependents were being discharged in higher proportion than was the case five months earlier. Among December separatees surveyed at five northeastern separation centers, slightly over 6 per cent expressed definite plans to attend full-time school. And, parenthetically, it may be stated that a War Department survey of enlisted men discharged in July 1946 found nearly 10 per cent expressing definite plans to attend full-time school.

Thus, it appears that there was considerable stability in soldiers' plans for full-time school, and that the fact of separation did not

markedly change men's plans in this respect. Therefore, it was relevant to inquire into the extent to which the expression of such plans at the time of discharge turned out to be predictive of efforts to attend full-time school.

Table 4 presents the classification of plans for full-time enrollment in school or college among the two sample groups of separatees who were followed up by mail to ascertain the extent to which they attempted to carry out their plans in the first two to four months after discharge. As earlier reported, data on their postseparation performance were secured from nearly 92 per cent of the July separatees and 88 per cent of the December separatees for whom legible

TABLE 4

PLANS FOR FULL-TIME SCHOOL ENROLLMENT EXPRESSED AT TIME OF DISCHARGE BY WHITE ENLISTED MEN

(July and December 1945)

(Per cent)

Plans for full-time enrollment	*July separatees*	*December separatees*
All separatees	*100.0*	*100.0*
Definite plans to enroll	3.1	5.7
Tentative plans to enroll	1.8	2.7
Considering enrolling	3.3	4.9
No expressed plans to enroll	91.8	86.7
Number of cases	*1,812*	*2,477*

addresses were available at the start of the follow-up. There was no educational bias in returns received from July separatees, but among December separatees there was a somewhat higher rate of return from high school graduates and college men than from those of lower educational attainment. In the data to be presented, all percentages are based on the total number of separatees actually contacted by the follow-up. Therefore, those percentages relating to plans of separatees may occasionally differ slightly from those reported for all separatees prior to the follow-up.

Postseparation Performance

Both for July and for December separatees, the number in school or college within three to four months after discharge was very close to the number who had expressed definite plans to attend school. But these were not all the men who could be expected to attend.

Three to four months was not a long enough interval to permit all men planning to enroll in school to carry out their plans. This was particularly true of the December dischargees, for whom an unforeseen contingency had developed. The limiting factor on their carrying out their plans was not the availability of government aid, but of space in the schools they wished to attend.

Many men were unable to enroll in the colleges they had planned to attend, largely because so many veterans were choosing the "name" colleges. Others had to postpone their enrollment for personal reasons. In addition, there were, of course, a number of men who had planned to take extended vacations before returning to school, and for these men, likewise, the fact that they had not enrolled within three or four months after discharge could not be considered a failure of prediction. Therefore, for purposes of evaluating the extent to which plans for school were predictive of actual returns, the criterion of successful prediction could not be solely the fact of enrollment.

Just as men could be ranked according to the plans for school which they expressed at the time of separation, they could also be ranked crudely according to the interest they had shown after discharge in enrolling or attempting to enroll in school. The following groups were distinguished in analyzing returns to the two follow-up surveys:

1. Men who reported they were actually enrolled in school full-time.
2. Men who reported that they definitely planned to enroll full-time and that they had already applied to the Veterans Administration for certificates of eligibility.
3. Other men who reported that they definitely planned to enroll full-time (many of whom had applied to one or more schools or colleges for admission).
4. Men who said they *might* enroll full-time.
5. All others, including those enrolled or planning to enroll part-time and those taking on-the-job training.[8]

Table 5 shows for July and December separatees the proportion in school or interested in attending at the time of the follow-up surveys, two to four months after discharge.

[8] It should be noted that veterans interested in education or training could receive aid for part-time schooling or for on-the-job training as well as for full-time schooling. In so far as veterans who had planned to attend full-time school shifted their interests to a part-time or job training course, they are here treated as complete errors in prediction.

The proportion of separatees who should be classified as having made a serious effort to attend full-time school probably lies somewhere between the number who were either already enrolled or who, planning definitely to enroll, had applied for aid, and the total number who said they were definitely planning to enroll (in other words, between 4.9 and 7.4 per cent for July separatees and between 6.9 and 10.0 per cent for December separatees).[9] Comments written in a number of the questionnaires of men who said they were planning to enroll but had not yet applied for certificates of entitlement indicated that the men were making an all-out effort to arrange for enrollment with the school before applying to the Veterans Administration for a certificate.

TABLE 5

POSTSEPARATION INTEREST IN FULL-TIME EDUCATION REPORTED BY SEPARATEES SURVEYED TWO TO FOUR MONTHS AFTER DISCHARGE

(Per cent)

Postseparation interest in full-time education	*July separatees*	*December separatees*
Total	*100.0*	*100.0*
Actually in school	3.0	4.7
Definitely planning to enroll	4.4	5.3
Had applied for aid	(1.9)	(2.2)
Had not applied for aid	(2.5)	(3.1)
Considering enrolling*	8.7	1.8
No expressed interest in full-time school	83.9	88.2

* July and December figures based on different criteria. July figure probably includes a good many men actually interested in part-time school

The number of men with definite plans for school had been regarded as the best estimate of the number who would actually attempt to return. Both follow-up surveys, however, indicate that the number who had returned or seemed likely to return as of three months after discharge was nearer to the total number who had expressed either definite or tentative plans for school.

It is apparent that in so far as the prediction problem is defined as estimation of the proportion of separatees who would seek to enroll in full-time school within a year of discharge from the service, only two observations are available to evaluate the use of plans for making such a prediction. Further, there is little to choose between

[9] This assumes that relatively few men had made an unsuccessful effort to enroll and had thereupon given up the idea of returning to school by the time of the follow-up.

the accuracy of the prediction for July separatees and that for December separatees. All that can be said is that in two instances where plans of separatees were used as a basis for estimating school enrollment of groups of separatees, the estimates seem to have been reasonably accurate if based upon the proportions with either definite or tentative plans to enroll.

Even at the time of final preparation of this volume, there is not yet available any definitive estimate of school attendance by veterans. As of December 31, 1946, the Veterans Administration reported approximately 1,680,000 veterans enrolled in institutional education under either Public Law 346, the Servicemen's Readjustment Act of 1944, or Public Law 16 providing educational benefits for disabled veterans. It may be assumed, from earlier sample surveys made by the Veterans Administration, that roughly seventy per cent of the veterans enrolled in schools and colleges are full-time students. Thus, one would obtain an estimate of 1,175,000 full-time students by the end of 1946, roughly six months after the completion of demobilization of the great bulk of the wartime armed forces. The figure is not markedly different from that estimated more than a year previously on the basis of the surveys conducted in the summer of 1945—"It seems likely that within six months to one year after total demobilization not less than one million men will have enrolled for full-time study in educational institutions." [10] It would appear, then, that so far as concerns the objective of estimating the proportion of veterans who would seek to enroll in school, surveys of soldiers' plans and intentions gave estimates reasonably accurate for general planning purposes under the conditions that actually prevailed during and subsequent to demobilization. It must be recognized that if a different set of circumstances had prevailed, the predictions might have been made less accurate. Had there been no increase in the amount of aid available to veterans in December 1945, the proportion enrolling in schools and colleges might have been considerably less—perhaps no more than the proportion who had expressed *definite* plans to enroll in school. On the other hand, if there had been eight million unemployed, as predicted by a number of leading economists for the immediate postdemobilization period, it is quite likely that the number of veterans seeking to enter schools and colleges would have been far greater. In any case, the limited facilities of our college system

[10] Information and Education Division, Report No. B-174, "Soldiers' Plans for Full-time School or College," December 28, 1945.

have undoubtedly held down to some extent the volume of enrollment which might have been attained if "name" colleges had been able to accept all qualified applicants.

But what of the accuracy of predictions made for the individual veteran on the basis of his plans? ·

Prediction of Enrollment for the Individual Separatees

Among respondents to the follow-up of July separatees, less than a third of those who had expressed definite plans to attend school had actually enrolled within the first few months after discharge and another fifth reported that they still planned to enroll. Men who had expressed only tentative plans to attend school apparently

TABLE 6

RELATIONSHIP OF PRESEPARATION PLANS FOR EDUCATION TO ENROLLMENT OR INTEREST IN FULL-TIME SCHOOL REPORTED TWO TO FOUR MONTHS AFTER SEPARATION

(July Separatees)
(Per cent)

POSTSEPARATION INTEREST IN FULL-TIME SCHOOL	PLANS FOR FULL-TIME ENROLLMENT AT TIME OF DISCHARGE			
	Definite plans	*Tentative plans*	*Considering enrolling*	*None expressed*
Total	*100*	*100*	*100*	*100*
In school	31	32	16	1
Definitely planning to enroll	21	26	9	3
Had applied for aid	(13)	(10)	(2)	(1)
Had not applied	(8)	(16)	(7)	(2)
Considering enrolling	15	29	20	8
No expressed interest	33	13	55	88
Number of cases	*52*	*31*	*56*	*1,509*

returned in about the same proportion as those who had been more definite. (The number of cases in each of these groups is so small as to make comparison exceedingly hazardous.) Further, about a fourth of the men who had indicated they were considering attending full-time school, and about 4 per cent of those who had not, at the time of separation, expressed any plans to attend, had either enrolled full-time or had become definite prospects for full-time education when surveyed two to four months after their discharge from the service (Table 6).

At the time the July (1945) separatees were returned to civilian life, there was still a great demand for workers in war industry. A number of the men indicated that they planned to take such a job during the period of emergency and then expected to return to full-time school. The war ended only a month later. It is not surprising that some of these men who had not predicted they would return to school immediately after discharge should actually have done so under the changed circumstances.

TABLE 7

RELATIONSHIP OF PRESEPARATION PLANS FOR EDUCATION TO ENROLLMENT OR INTEREST IN FULL-TIME SCHOOL REPORTED THREE TO FOUR MONTHS AFTER SEPARATION

(December Separatees)

(Per cent)

POSTSEPARATION INTEREST IN FULL-TIME SCHOOL	PLANS FOR FULL-TIME ENROLLMENT AT TIME OF DISCHARGE			
	Definite plans	*Tentative plans*	*Considering enrolling*	*None expressed*
Total	*100*	*100*	*100*	*100*
In school	46	16	8	1
Definitely planning to enroll	32	36	21	2
Had applied for aid	(14)	(18)	(6)	*
Had not applied	(18)	(18)	(15)	(2)
Considering enrolling	2	8	4	2
No expressed interest	20	40	67	95
Number of cases	*125*	*62*	*103*	*1,808*

* Less than .5 per cent

The follow-up of December (1945) separatees showed a somewhat closer relationship between plans for education and postseparation performance. It will be seen (Table 7) that a clear gradation exists in the extent of returns to school by plans classification—definite, tentative, etc. Not only did a higher proportion of the men with definite plans tend to carry out those plans but a lower proportion of those who had *not* planned to return to school actually did return.

Tables 6 and 7 indicate that, although a certain proportion of men who had not expressed at time of discharge either definite or tentative plans to enroll in school actually did enroll, in general one would predict enrollment only for those individuals who had expressed definite or tentative plans. The relationship between these cate-

gories of soldiers' plans and postseparation performance is summarized in Table 8. The definition of performance was to be "an attempt to enroll in school." Two criteria of performance are used in Table 8: Criterion A—men who reported themselves either in school or as planning to enroll full-time and as having applied for educational benefits; Criterion B—same as A plus all other men who reported definite plans to enroll full-time when surveyed after separation. In general, it is felt that Criterion B more nearly represented the group likely to enroll or make an effort to enroll during the first six months to a year after discharge.

TABLE 8

PLANS FOR FULL-TIME SCHOOL RELATIVE TO TWO CRITERIA OF ENROLLMENT

(July and December Separatees)

(Per cent)

PLANS TO ATTEND FULL-TIME SCHOOL	POSTSEPARATION PERFORMANCE					
	Criterion A			*Criterion B*		
	Total	Enrollees*	Others	Total	Enrollees**	Others
July Separatees						
Total	*100.0*	*4.9*	*95.1*	*100.0*	*7.4*	*92.6*
Definite	3.1	1.4	1.7	3.1	1.6	1.5
Tentative	1.8	0.8	1.0	1.8	1.1	0.7
All others	95.1	2.7	92.4	95.1	4.7	90.4
December Separatees						
Total	*100.0*	*6.9*	*93.1*	*100.0*	*10.0*	*90.0*
Definite	5.9	3.6	2.3	5.9	4.6	1.3
Tentative	3.0	1.0	2.0	3.0	1.5	1.5
All others	91.1	2.3	88.8	91.1	3.9	87.2

* Includes those actually enrolled and those planning to enroll full time who had applied for aid.
** Includes those actually enrolled and all those planning to enroll full time.

Table 8 clearly indicates that in two respects the plans expressed by December separatees were more highly predictive of postseparation performance than were plans of July separatees: (1) among those who expressed plans to go to school, a higher proportion carried out or sustained such plans; (2) among those who actually enrolled or seemed likely to enroll on the basis of follow-up data, a higher proportion consisted of individuals who had reported plans to return to school.

It is of interest to inquire into the reasons for the greater degree of association between plans and performance of December separatees as compared with July separatees, and also to inquire in gen-

eral into the reasons why individual predictions based on soldiers' plans to attend full-time school were not more often correct. In doing so, it is convenient to consider three aspects of the prediction problem:

1. The adequacy of the classification of soldiers' plans to attend full-time school.
2. The characteristics of the members of the population surveyed.
3. The situation encountered after discharge.

Each of these aspects will be briefly discussed in the pages that follow.

Adequacy of plans classification. Evaluation of the adequacy of the classification of soldiers' plans used as a basis for the prediction of enrollment for full-time study in school or college involves a reconsideration of the structuring of the prediction problem, an examination of the predictive value of specific items used in the questionnaire and of the method whereby such items were combined and, finally, consideration of alternative approaches.

It will be recalled that in the major surveys of soldiers' plans, classification of plans for education was based upon responses to three items which had been used to define major scale types in earlier pretests and surveys. The selection of only three items as a basis for classification sacrificed one of the chief advantages of the use of a scale—that of permitting ready classification of respondents who failed to answer one or two questions in a battery or answered one or two questions in such a way as to suggest misinterpretation or erroneous checking. On the other hand, the use of a long battery of questions on educational plans in a questionnaire relating to a large number of fields which veterans might wish to enter was not deemed feasible. Indeed, a long battery of questions which more or less repeated the same inquiry might be expected to produce more erroneous checks and failures to answer. And since the items selected were items which produced the scale pattern with relatively low degree of error, categorization based in part on consistency of response (rather than being based entirely on scaling to type) did not markedly reduce the size of the groups classified as having definite or tentative plans to attend full-time school. At the same time, the experience gained with the longer set of items used in the earlier survey of educational plans and in pretests suggested those combinations of responses that should be made in attempting to arrive at a small number of categories for the classification of those men somewhat interested in further education but not really planning to be-

come full-time students after discharge. How important were these categories in their contribution of veteran students?

Tables 6 and 7 indicated that the proportion of veterans who returned to full-time school or were considering returning at the time of the follow-up, from among all separatees who had not expressed any intentions of becoming full-time students, was small—4 per cent among July separatees and 3 per cent among December separatees. But the contribution to total enrollment of veterans made by separatees who had not planned to return to school was nevertheless considerable because of the magnitude of this segment of the population relative to the number who had planned to enroll. Table 9 shows in detail the contributions made to the total full-time and part-time enrollment or prospective enrollment of December separatees by the various categories of educational plans expressed at the time of discharge. It is apparent that in the case of both full-time and part-time enrollment relatively few students, or potential students, were drawn from that half of the sample which had expressed no interest in education at the time of discharge. Of full-time enrollees, in fact, three fourths were contributed by the 14 per cent of the sample who had indicated that they were either planning

TABLE 9

DISTRIBUTION OF TOTAL SAMPLE AND OF VETERANS ENROLLED OR PLANNING TO ENROLL IN FULL- OR PART-TIME SCHOOL, BY EDUCATIONAL PLANS EXPRESSED AT DISCHARGE

(December 1945 Separatees)
(Per cent)

PLANS EXPRESSED AT TIME OF DISCHARGE	TOTAL SAMPLE	FULL-TIME STUDENTS		PART-TIME STUDENTS
		In school	*Planning to enroll*	*In school or planning to enroll*
Total	*100*	*100*	*100*	*100*
Full-time school	14	76	75	19
Definite plans	(6)	(58)	(35)	(6)
Tentative plans	(3)	(10)	(20)	(5)
Considering enrolling full-time	(5)	(8)	(20)	(8)
Plan to enroll part-time	11	6	12	46
Vague interest in education but no plans indicated	26	9	7	19
No expressed interest	49	9	6	16
Number of cases	*2,098*	*99*	*112*	*118*

to attend school full-time or at least were considering full-time enrollment at the time of discharge.[11] Nevertheless, it is pertinent to inquire into the general plans classification of those men who were categorized as having plans for part-time school or as having merely vague interest, or even no expressed interest, in school yet who reported themselves in school or planning to enroll some three or four months after discharge. Were these men undecided as to their plans at the time of discharge, or did they have other plans which were modified by the experiences of the first few weeks of civilian life? The latter seems very largely to have been the case. In answer to questions involving alternative plans, these men had indicated in large proportion that they planned to work as employees. Some had expected to return to work for the employer they had been with before the war; a larger number had expected to find a new job. Unfortunately, the number of cases involved is too few to permit any adequate analysis of the reasons for the shift which occurred in the plans of these men, or of their characteristics.

What estimates as to probable enrollment of veterans as full-time students would have been made if responses to any one of the individual questions relating to plans for school had been taken as the basis for plans classification?

Let us consider first the two questions which served as the basis for establishing the cutting points which defined the group described as having definite plans and that described as having tentative plans —the question involving choice of full-time school among several alternatives (Q 22) and the direct question relating solely to plans for school attendance (Q 39). It is to be noted that responses to the separate items will give different estimates from those derived from the use of the questions in a scale in so far as the items show other than a random distribution of scale error. Since the three questions used for a basic classification of separatees' school plans were too few to permit scoring nonperfect types in accordance with scale theory, complete consistency of response had to be the criterion of classification as having definite plans. It will be apparent, therefore, that the question involving choice of alternatives, used to define the type called definite planners, actually yielded a higher estimate of school enrollment when used alone than when used in

[11] It is noteworthy that although nearly half of the part-time enrollees were drawn from the 11 per cent of the December sample who expressed plans for part-time school, more than a third of the part-time enrollees come from the three fourths of the sample who expressed little or no interest in education at the time of separation. The problem of predicting part-time education will be discussed briefly later in this section.

conjunction with other questions (since those not consistent in their expressed plans would be eliminated). The estimates obtained from this source, and their relationship to performance, are shown in Table 10.[12] From a comparison of Table 10 with Table 8 (using Criterion B), it will be noted that a prediction based upon question 22 would be higher than that yielded by the classification "definite plans" but slightly lower than the combined "definite" plus "tentative" plans categories. Accepting Criterion B as the best estimate of actual enrollment after discharge, it would appear that the higher prediction given by the combined "definite" and "tentative" categories using three questions would be a closer approach to the mark than the prediction based on question 22 alone. It is to be noted, moreover, that when the "definite" and "tentative" categories of plans in Table 8 are combined, not only is a higher prediction of attendance given but a higher proportion of the group desig-

TABLE 10

RELATIONSHIP OF RESPONSES TO INITIAL PRESENTATION OF ALTERNATIVES IN SURVEY OF PLANS AND POSTSEPARATION SCHOOL ENROLLMENT

Criterion B—July and December Separatees)

(Per cent)

QUESTION 22. Regardless of whether you have definite plans or not, which one of these things do you think you will most probably do FIRST—right after you leave the Army?

RESPONSE CATEGORY CHECKED	STATUS AT TIME OF FOLLOW-UP: *July separatees* Total	In school or planning to enroll	All others	*December separatees* Total	In school or planning to enroll	All others
Total	*100.0*	*7.4*	*92.6*	*100.0*	*10.0*	*90.0*
Work for an employer*	85.4	5.3	80.1	82.2	4.4	77.8
Go to full-time school or college	4.7	2.1	2.6	8.1	5 3	2.8
Run own business or farm	8.4	—	8.4	7.6	0.2	7.4
No answer	1.5	—	1.5	2.1	0.1	2.0

* Includes those who planned to work in family businesses or on family farms For exact wording of response categories, see Appendix A to Chapter 16.

[12] The question wording used in the surveys of separatees will be seen to differ somewhat from that used in the earlier surveys, but the fundamental relationships of the questions themselves were unchanged. The question numbers reported are from the survey of separatees and will serve as guidance to the reader who wishes to consult the questionnaire itself, contained in Appendix A to Chapter 16.

nated as likely to enroll actually does so. Thus, of the 4.7 per cent of July separatees who chose the alternative "go to full-time school or college" in answer to question 22, 2.1 or less than half (45 per cent) were in school or planning to enroll as of three to four months after discharge, but of the 4.9 per cent classified as having definite or tentative plans by the modified scale, 2.7 or slightly over half (55 per cent) were in school or seemed likely to enroll. Again, among December separatees the choice of school in question 22 by 8.1 per cent of the group was a valid prediction of attendance for about 65 per cent of those who made the choice; but of the 8.9 per cent classified as having definite or tentative plans, 69 per cent actually qualified as students by the same criteria.[13]

As a basis for prediction of individual enrollment, then, the question involving a choice of alternatives would seem to be less satisfactory than its use in combination with the other scale items used.

Reference to the role of the direct check-list question on plans for school (Q 39) in the classification of plans indicates that it gave the cutting point between those men classified as having definite or tentative plans to attend school and all others. Table 11 indicates that the use of this question alone differs hardly at all from the use of the three items selected to define scale types if the types are dichotomized into "definite or tentative planners" versus "all others."[14] It should not be inferred, however, that the single direct question provides an adequate basis for classifying soldiers' plans for school. Reference to Table 7 clearly indicates that December separatees classified as having definite plans to attend school actually enrolled in considerably higher proportion than did those with only tentative plans. If conditions had been less favorable to school enrollment of veterans, it might well have been that few except those classified as having definite plans actually would have enrolled.

It had been hoped that by inserting at the close of the questionnaire a second opportunity to indicate choice among alternatives it might be possible to get a more reasoned choice and hence a better basis for prediction than would be provided by the earlier question involving alternatives (Q 22). Such did not prove to be the case,

---

[13] The differences are not statistically significant if the high degree of correlation between membership in the two categories is ignored, but that for July separatees exceeds the 2σ level if one assumes a .5 correlation.

[14] The source of the small difference that does occur is the fact that a few of the men who said they planned to go to full-time school in answer to question 39 had not indicated that they *wanted* to go to full-time school in answer to question 38.

however. For each of the samples of separatees, the number of men checking that they planned to attend school was slightly less for the second question involving alternatives than for the first, but the predictive value of the choices was not appreciably greater, as will be seen by comparison of the data given in Table 12 with the previously reported data for question 22 (Table 10). Table 12 also gives some detail on the plans of men who were at the time of

TABLE 11

RELATIONSHIP OF RESPONSES TO DIRECT QUESTION ON SCHOOL PLANS AND POSTSEPARATION SCHOOL ENROLLMENT

(July and December Separatees)

(Per cent)

QUESTION 39: Regardless of what you would like to do, do you think you will actually go back to school or college after you leave the Army?

	STATUS AT TIME OF FOLLOW-UP					
	July separatees			*December separatees*		
RESPONSE CATEGORY CHECKED	Total	In school or planning to enroll	All others	Total	In school or planning to enroll	All others
Total	*100.0*	*7.4*	*92.6*	*100.0*	*10.0*	*90.0*
Yes, to *full-time* school or college	5.4	2.8	2.6	9.1	6.2	2.9
Yes, to *part-time* school or college	14.5	2.0	12.5	13.6	1.4	12.2
No, I don't think I will go back to school or college	52.4	1.0	51.4	47.5	0.6	46.9
Undecided	20.5	1.4	19.1	22.9	1.6	21.3
No answer	7.2	0.2	7.0	6.9	0.2	6.7

the follow-up either enrolled in school or planning to enroll although they had not at the time of separation expressed plans to enroll.

By following the final choice of alternatives with a question asking for degree of certainty of plans, it was possible to distinguish between those men who felt quite sure they would enroll in school and those who, while leaning toward school enrollment, were somewhat less sure.[15] Table 13 shows for December separatees the extent to which feeling of certainty was related to the carrying out of plans.

[15] See question 69 in the questionnaire for separatees appended to Chapter 16.

TABLE 12

RELATIONSHIP OF RESPONSES TO FINAL PRESENTATION OF ALTERNATIVES IN SURVEY OF PLANS AND POSTSEPARATION SCHOOL ENROLLMENT

(July and December Separatees)

(Per cent)

QUESTION 68: In this questionnaire we have asked you a lot of questions about your plans for after your discharge. Now that you have thought them over a little more, try to take everything into consideration and check the *one* thing you will most probably do *first of all*—after you have gotten your discharge and taken a vacation.

RESPONSE CATEGORY CHECKED*	STATUS AT TIME OF FOLLOW-UP: *July separatees* Total	*July separatees* In school or planning to enroll	*July separatees* All others	*December separatees* Total	*December separatees* In school or planning to enroll	*December separatees* All others
Total	*100.0*	*7.4*	*92.6*	*100.0*	*10.0*	*90.0*
Work for pre-service employer	25.6	0.3	25.3	46.8	1.6	45.2
Work for some other employer	48 7	3.9	44.8	24.7	2.0	22.7
Go to full-time school	4.7	2.4	2.3	7.4	4.9	2.5
Go into business for self	5.4	0.1	5.3	5.6	0.2	5.4
Farm	5.2	0.1	5.1	3.3	0.2	3.1
Undecided	6.0	0.4	5.6	6.5	0.7	5.8
No answer	4.4	0.2	4.2	5.7	0.4	5.3

* For exact wording of response categories, see Appendix A to Chapter 16.

TABLE 13

CERTAINTY OF PLANS EXPRESSED BY CHOICE AMONG ALTERNATIVES (QUESTION 68) AND POSTSEPARATION ENROLLMENT IN FULL-TIME SCHOOL

(December Separatees)

(Per cent)

PLANS CLASSIFICATION	STATUS AT TIME OF FOLLOW-UP: *Total*	*In school or planning to enroll*	*All others*
Total	***100.0***	***10.0***	***90.0***
Full-time school	**7.4**	**4.9**	**2.5**
Sure of plans	**(4.5)**	**(3.7)**	**(0.8)**
Less sure	**(2.9)**	**(1.2)**	**(1.7)**
Alternative plans	**92.6**	**5.1**	**87.5**
Sure of plans	**(43.6)**	**(2.1)**	**(41.5)**
Less sure	**(49.0)**	**(3.0)**	**(46.0)**

Of men who had indicated that they felt sure of their plans to enroll in school after discharge, 84 per cent actually enrolled or seemed likely to enroll; of those who felt less sure, only 42 per cent had enrolled or sustained their plans to enroll. Here, then, was a method of further subdividing the group classified by the other items used as having definite plans, and a method which did give considerable discrimination judged by actual performance. In so far as interest is focused upon accurate individual prediction, the combination of choice of alternatives and degree of certainty of choice definitely yields a gain. It will be apparent, however, that with conditions favorable to school enrollment by veterans, the choice of alternatives alone seriously understates the prospective returns to school by classifying under other rubrics those men seriously considering school enrollment but leaning somewhat more strongly to some other plan at the time of discharge. Further discussion of the general structuring of the prediction problem and the consideration of alternative approaches will be deferred until after a brief analysis of the relationship of personal characteristics and degree of carrying out plans for school attendance.

Population characteristics and prediction of enrollment. The great majority of soldiers who expressed plans to attend school after discharge were young single men who had been out of school a relatively short time when they were inducted into the service. Even among men discharged from the Army in December 1945, although single men under twenty-five years of age and out of high school a year or less before induction numbered only 10 per cent, they contributed 56 per cent of the group with definite plans to attend school and 31 per cent of those with tentative plans.

A question much discussed in attempting to arrive at estimates of veteran enrollment was whether older married men and those longer out of school at the time of discharge were at all likely to return to school, even if they did have plans to go. It might be expected that those lacking the usual characteristics of a school population would run into more problems and difficulties in attempting to become students than would the young single group differing only by a few years of age from the normal college population. Such reasoning apparently resulted in limiting educational benefits of the original Servicemen's Readjustment Act of 1944 to veterans under twenty-five and those whose education was interrupted by the war. On the other hand, it seemed possible that the older or married men and those longer out of school might well be more

strongly motivated toward further education if they formulated plans for enrollment in the face of the difficulties to be anticipated.

Examination of the results of the follow-up of December separatees indicates that, among the older or married men and those out of school more than a year before they entered the service, the proportion carrying out or sustaining their plans in the first few months after discharge was less than that of the younger single group more recently from school when inducted. It will be seen in Table 14 that for each category of school plans, the group having the characteristics usually associated with student status was more likely

TABLE 14

POSTSEPARATION RETURNS TO SCHOOL BY PLANS EXPRESSED AT SEPARATION AMONG SINGLE MEN UNDER AGE TWENTY-FIVE OUT OF SCHOOL A YEAR OR LESS, AND AMONG OTHER SEPARATEES

(December 1945 Separatees)
(Per cent)

STATUS THREE TO FOUR MONTHS AFTER DISCHARGE	SINGLE MEN UNDER 25, OUT OF SCHOOL ONE YEAR OR LESS				ALL OTHERS			
	Total	*Definite plans*	*Tentative plans*	*Other*	*Total*	*Definite plans*	*Tentative plans*	*Other*
Total	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*
In school or college	25	54	21	8	2	35	14	1
Planning to enroll	20	29	32	13	4	36	35	2
Had applied for aid	(9)	(16)	(16)	(3)	(2)	(13)	(19)	(1)
Had not applied	(11)	(13)	(16)	(10)	(2)	(23)	(16)	(1)
Not planning to enroll	55	17	47	79	94	29	51	97
Number of cases	*211*	*70*	*19*	*122*	*1,887*	*55*	*43*	*1,789*
Percentage constituted among all veterans in school or planning to enroll	45	28	5	12	55	18	10	27

to return to school. Nevertheless, because they comprised so large a proportion of the total sample, the group less likely to return did contribute a majority of the students from among December separatees. (The data are contained in the bottom line of Table 14.)

One reason for the higher predictive value of plans expressed by the December separatees as compared with July separatees seems to have been the differential proportion of men having the characteristics usually associated with school performance. Among the July separatees only about 5 per cent were single, under twenty-five, and out of school less than a year before induction, and of these, a

considerable number had entered the Army during peacetime so that their education had not in any sense been interrupted by the war.

In summary, it appears that plans for postseparation school enrollment were more highly predictive of enrollment of veterans having characteristics most similar to those of a school population than of enrollment of older veterans, married veterans, or those who had left school more than a year before they entered the service. On the other hand, among the younger, unmarried veterans out of school only a short time before they entered the service, those who had not planned to return to school after discharge were more likely to shift in the direction of enrolling, so that the overall accuracy of prediction from plans alone was not appreciably different from that obtaining for the older men, those married and those longer out of school. It would appear that those having the characteristics most similar to a school population were, upon discharge, most likely to encounter a situation which they defined as favorable to enrollment.

The situation encountered after discharge. It was earlier observed that implicit in the plans expressed by soldiers and separatees was an evaluation of various contingencies relating to the situation to be encountered after discharge. Plans expressed in the summer of 1944 on the basis of an assessment of future contingencies as they appeared at that date might readily be shifted as a result of a continuing reassessment of those contingencies. Specifically, the man whose plans for school in the summer of 1944 rested on the conviction that he would be demobilized within a year and could start college in the fall of 1945 might well decide to go to work upon learning a year later that he would be in service well into 1946. This would be particularly true if in that intervening year he acquired a wife or a prospective wife.

The surveys of separatees indicated that there had not been any major change from July 1944 to December 1945 in the proportion of soldiers planning to enroll in school, although there had been a slight increase. Nevertheless it must be assumed that a number of soldiers decided not to attend and a slightly larger number decided to attend school as a result of changed evaluations of the situation to be encountered after discharge. Since the follow-up studies were conducted only among separatees, they do not afford adequate data on the magnitude of such changes. Among men discharged in July 1945, however, there is a clear indication that within two months after discharge a considerable number of veterans reevalu-

ated the opportunities available to them and shifted their plans accordingly. When they first learned that they would be among the first men demobilized, many veterans discharged at this time resolved to work in war industry, in order to get a chance at some of the high wages they had heard so much about while in service, and others decided to seek a good permanent job while the labor market still made great demands for workers. With the surrender of Japan, however, a number of these men decided to go to school instead. It is quite possible that such men would have expressed plans for school if surveyed a year earlier, and that the termination of the war thus resulted in a return to plans previously modified by prospects of war industry employment. The fact remains that the changed situation seems to have resulted in a more serious underestimate of school enrollment of July separatees than was true of December separatees.

It has already been mentioned that the increase in allowances to veterans attending schools, provided by amendment in December 1945 to the earlier Readjustment Act, probably increased slightly the number of veterans actually enrolling in schools and colleges, and the fact that many veterans had inflated notions of the sort of jobs available to them may also, after deflation of such notions, have resulted in some increase in enrollment. Countering this, crowding in the colleges, which absorbed about two thirds of the full-time enrollees, has probably limited somewhat the proportion of later separatees who were able to carry out their plans to enroll in school.

In concluding this review of the attempt to predict the enrollment of veterans in schools and colleges from a knowledge of plans expressed while in service, brief consideration is given to the relationship between plans to enroll in school on a part-time basis, and actual part-time enrollment within a few months after discharge. Slightly over 11 per cent of the December separatees had reported that they would like to attend part-time school and that they thought they actually would attend after discharge. As has already been indicated (Table 9), the men planning to attend part-time school comprised nearly half of those who were enrolled or expected to enroll at the time of the follow-up. But, as will be seen in Table 15, less than a fourth of the men who had expressed plans for part-time school had actually carried out or sustained their plans. Further, those who had not carried out plans for part-time school were not counterbalanced by any substantial enrollment by men who had

not expressed plans to enroll. This situation had early been foreseen in reporting soldiers' plans. The fact that plans for part-time school might be compatible with almost any other activity meant that a considerable number of men might report their own wishful thinking. There was nothing in the questionnaire to force the soldier to evaluate his own plans for part-time school.

Table 15 also indicates that of those veterans who did not sustain their plans for part-time school, an appreciable proportion shifted either to full-time school or to on-the-job training. Thus, it is to

TABLE 15

RELATIONSHIP BETWEEN VARIOUS CATEGORIES OF SOLDIERS' EDUCATIONAL PLANS AND EDUCATIONAL OR JOB-TRAINING STATUS THREE TO FOUR MONTHS AFTER DISCHARGE

(December 1945 Separatees)
(Per cent)

EDUCATIONAL STATUS AT TIME OF FOLLOW-UP	PLANS FOR EDUCATION EXPRESSED AT DISCHARGE				
	Total	*Full-time school**	*Part-time school*	*Vague interest, no plans***	*No interest*
Total	*100.0*	*13.8*	*11.4*	*25.5*	*49.3*
Full-time school	10.0	7.6	0.9	0.8	0.7
In school	(4.7)	(3.6)	(0.3)	(0.4)	(0.4)
Planning to enroll	(5.3)	(4.0)	(0.6)	(0.4)	(0.3)
Part-time school	5.5	1.1	2.6	1.0	0.8
In school	(1.2)	(0.5)	(0.5)	(0.1)	(0.1)
Planning to enroll	(4.3)	(0.6)	(2.1)	(0.9)	(0.7)
On-the-job training	11.6	1.2	1.9	3.3	5.2
In training	(4.6)	(0.4)	(0.8)	(1.0)	(2.4)
Planning to enter	(7.0)	(0.8)	(1.1)	(2.3)	(2.8)
All others	72.9	3.9	6.0	20.4	42.6

* Includes those with definite and tentative plans plus those considering enrolling full-time.
** Includes those who said they would like to enroll but did not plan to do so.

be noted that either an intensified interest in education, leading to full-time enrollment, or a reevaluation of training opportunities, leading to a choice of on-the-job training, would result in wrong prediction just as would a decline in interest leading to a discarding of all plans for education.

To soldiers interested in further education there was available aid for full-time school, for part-time school, or for on-the-job training. Full-time school was most often chosen by young single men, particularly by those who had been in school just before they entered the service. Part-time school appealed more often to the married men who felt they could not afford to return to school full-time,

and to single men who had been working before entering the service. Relatively few soldiers knew about the possibilities of on-the-job training until well along in the demobilization period.[16] While the types of training for which aid was available are in some respects quite dissimilar, they did apparently serve in some instances as adequate substitutions, one for another, so that no real change of plans would be involved. This would be particularly true in the case of on-the-job training as a substitute for business or trade school, and part-time college as a substitute for full-time enrollment.

One further note is relevant: the 1944 cross-sectional survey found that of soldiers definitely planning to enroll in full-time school, two thirds planned to enroll in college, about a fourth planned to enroll in trade and business school courses and less than 5 per cent planned to complete standard high school courses. Veterans Administration operating statistics as of summer 1946 showed veterans enrolled in educational institutions to have almost identical distribution by type of school with that predicted on the basis of soldiers' plans.

[16] As of July 30, 1945, Veterans Administration statistics showed 20,229 veterans in institutional education and 2,120 in on-the-job training. By July 30, 1946, the number reported in institutional education had increased thirtyfold (to 652,762), but the number reported in on-the-job training had increased two hundredfold (to 403,199).

CHAPTER 16

STUDIES OF THE POSTWAR PLANS OF SOLDIERS—CONTINUED[1]

SECTION I

PLANS TO RETURN TO PREVIOUS EMPLOYER

APPROXIMATELY four fifths of the Army's male personnel had worked for an employer just before entering the service. The Selective Service Act of 1940 assured reemployment rights to all personnel entering the Armed Forces after May 1, 1940, who were occupying nontemporary jobs in the civilian economy as of the time of induction or enlistment. Some men, however, had left their regular jobs either to take jobs in war industry or to take vacations before entering the service. Once the war was under way, many men entering war industry were replacing others who had enlisted or been inducted.

There was no accurate knowledge of the number of men in service who had reemployment rights to their preservice jobs. Estimates generally ran between 20 per cent and 30 per cent and were based largely on guesswork, as the following quotation suggests:

"When allowance is made for those veterans, now probably 40 per cent of the total, who entered military and naval service without previously having had gainful employment, for the large number who will have no reemployment rights because they occupied temporary positions or were unemployed at the time of induction, for those whose jobs will be wiped out by business failures or technological changes, for those so disabled that they will not be able to resume their old occupations, and for those jolted out of a customary way of life, whom new maturity, skills, or ambitions will tempt to greener pastures, the number of veterans with effective rights to reinstatement is estimated at about 20 per cent of the total." [2]

[1] By John A. Clausen.

[2] Walter J. Couper, "The Reemployment Rights of Veterans," *Annals of the American Academy of Political and Social Science*, Vol. 238 (March 1945), p. 112.

The estimate that 40 per cent of all military personnel had not previously been gainfully employed was at least double the actual figure, but there was no way of knowing how many veterans had occupied temporary jobs or how many desired to find new jobs. Moreover, many employers indicated that they would attempt to reemploy not merely those veterans who had legal rights to a job, but all veterans who had been in their employ at the time of induction or enlistment. It seemed relevant, therefore, to inquire into the number of men in service who planned to return to work for their previous employers.

Ascertaining Plans to Return to Previous Employer

The problem of multidimensionality of plans was nowhere more difficult to deal with than in the area of plans to return to work for one's previous employer. Some men wanted to return to the exact niche they had previously occupied; others wanted to return to the same employer but in a different (and usually a better-paying) job; still others were interested in working in the same occupation but with a different employer, yet were considering the possibility of going back to work for their former employer if other opportunities were not readily available. For a large proportion of the Army's personnel who had previously been gainfully employed, it appeared that the possibility of returning to work, at least temporarily, for one's previous employer was not to be completely dismissed even though other pastures might seem greener. A survey conducted in December 1943 to elicit attitudes toward postwar vocational guidance included three questions relating to interest in doing the same or a different kind of work with the same or a different employer after the war. More specifically, the questions asked were:

1. Whether the respondent *would like to go back* to his old job with his old employer, a different job with his old employer, the same kind of work with a different employer, or a different kind of work with a different employer.
2. What kind of work the respondent *would like to do* after the war.
3. What kind of work the respondent *actually expected to do* after the war.

It was found that of men who had previously been employed, a slight majority (54 per cent) said they would like to do the same kind of work, and of this group, about three fourths said they would like to work for their previous employer. Among the 46 per cent of all previous employees who said they would like to do a different

kind of work, however, nearly half were doubtful as to whether they would actually do a different kind of work, and a fourth thought they would probably return to work for their previous employer. Clearly this was not a unidimensional problem.

Considerable difficulty was encountered in attempting to phrase questions simply enough to be intelligible to the soldier with only fifth or sixth grade education yet precisely enough to specify the dimensions desired for classification. For example, the expression "go back to your old job after the war," was interpreted by a majority of soldiers as referring to a job with their last employer prior to induction, but some men took this expression to mean the same general occupation, regardless of employer, and some took it to mean precisely the same occupation with the same employer. The man who hoped to go back to a somewhat different job with his previous employer was frequently at a loss to know how he should classify himself. To get around this problem, it was decided to ask separate questions about type of work or occupation, and about prospective employer.

Another major difficulty was encountered in inconsistencies among responses to questions which were designed primarily for men who had been employed just before their induction. Such questions usually carried an answer category, "I was *not* working just before I entered the Army," but some men who had not previously been employed failed to check this category and instead checked that they planned to do a different kind of work than they had previously done. Other men, not employed just before induction but employed at some earlier date, further confused the picture with respect to probable reemployment of those drafted or volunteering for the service from their jobs subsequent to the spring of 1940. To meet this problem it seemed desirable first of all to ascertain employment status prior to induction and then in this area to study the plans only of those men who had occupied full-time jobs prior to their induction.

In the world-wide survey conducted in the summer of 1944, men were classified first as to whether or not they had been employed just prior to induction or enlistment, then by their judgment of whether they could get work with their previous employer and, finally, whether they expected actually to return.[3] They were also

[3] The questions asked on these topics were as follows:
After the war, do you think you could get work with the company or person you worked for before you came into the Army, if you wanted to? (Check one)

asked to describe the kind of work they expected to do, and this was coded as "similar to" or "different from" the kind they had previously done. These data were used as a basis for estimating the minimum number of soldiers counting on going back to their old jobs and the maximum number considering returning to work for their previous employers. The method of classification is described in the following quotation from the survey report:

"Those enlisted men who say both that they plan to do the same kind of work and that they *expect* to return to their old jobs may be considered a minimum estimate of the proportion definitely counting on going back to their old jobs. These comprise a fourth of the white enlisted men who worked as employees before they entered the Army. Those men (fourteen per cent of former employees) who are planning to return to the same kind of work and who say they *may* return to their former employers also seem more likely to try to get their old jobs back than to make a change. The estimate of the maximum number of men who are considering returning to their former employers includes all other men (twenty-one per cent) who say that they *may* go back to work for the employer they worked for just before they entered the Army. Some of these men are undecided about their postwar plans, and many of them are seriously considering alternative plans—a different type of work for an employer, full-time school, self-employment, etc.—but they do not feel sure enough about the alternatives to rule out going back." [4]

It appeared likely then, on the basis of their plans, that about 40 per cent of the soldiers who had previously been employed might return to work for their former employers. There was some doubt, however, as to the desirability of using the dimension "type of work"

______ Yes, I'm almost sure I could
______ Yes, I think so, but I'm not sure
______ No, I probably couldn't
______ I'm sure I could not

______ I worked for myself before I came into the Army
______ I was not working before I came into the Army

Do you think you *actually will* go back to work for the same employer (company, person, etc.) you worked for before you came into the Army? (Check one)

______ Yes, I'm quite sure I will
______ I may, but I'm not sure
______ No, I don't think I will

______ I worked for myself before I came into the Army
______ I was not working before I came into the Army

[4] Information and Education Division Report No. B–129, *Post-War Occupational Plans of Soldiers*, March 1, 1945, p. 18.

as a criterion for estimating probable extent of return to previous employer. A man might desire a change in occupational category yet primarily be intent on securing a job with his old employer. Subsequently, therefore, the use of this dimension was discontinued and a different criterion for distinguishing between probable returnees and others was established.

Plans of Separatees in July and December 1945

The questionnaire administered to separatees used the same basic questions for a classification of plans to return to preinduction employer with but one modification of any importance: in answer to the question, "Do you think you actually will go back to work for the same employer . . .?" the previously used response category, "I may, but I'm not sure" was replaced by two new categories: "Yes, I probably will" and "I may, but I probably will *not*." It was hoped that under conditions favorable to prediction from stated plans, this change would permit discrimination between one group of which more than half the members would actually return and another group of which somewhat less than half would return. On the assumption that errors of individual prediction would to a considerable extent counterbalance each other, the new categories gave a cutting point for a total percentage prediction as well as for individual prediction. The classification scheme, along with the two questions on which it was based, is given below.

Question 24: Do you think you could get work with the employer you worked for before you came into the Army, if you wanted to?

Question 25: Do you think you *actually will* go back to work for the same employer you worked for before you came into the Army?

Classification	*Answer to question 24*	*Answer to question 25*
Definite plans to return	Yes, I'm almost sure I could	Yes, I'm almost sure I will
Tentative plans to return	Yes, I'm almost sure I could	Yes, I probably will
	or	
	Yes, I think so, but I'm not sure	Yes, I'm almost sure Yes, I probably will
Considering returning but not planning to	Yes, I'm almost sure I could Yes, I think so, but I'm not sure	I may, but I probably will *not*
Not considering returning	All other combinations	

It was found that among July separatees at least one sixth of the men had entered the Army before the Selective Service System came into effect. These men, of course, tended to have the highest accumulation of points for demobilization. Somewhat less than two thirds of the July separatees had been gainfully employed at full-time jobs just prior to entering the service. Another 11 per cent reported that they had had part-time jobs before entering the service. Among December separatees, on the other hand, slightly over four fifths had been gainfully employed at full-time jobs and about 3 per cent had been employed part-time before entering the service. The preinduction employment experience of the December separatees approximated that of the total Army as revealed by earlier world-wide surveys. This was in accordance with expectation, since they were being released about midway in the demobilization process and in most characteristics were fairly close to Army averages as of peak strength.

TABLE 1

COMPARISON OF PLANS OF WHITE ENLISTED MEN TO RETURN TO PREVIOUS EMPLOYER AMONG MEN IN SERVICE IN SUMMER 1944 AND MEN SEPARATED FROM SERVICE IN JULY AND DECEMBER 1945

(Per cent)

<table>
<tr><th></th><th>Men in service Summer 1944</th><th>Separatees July 1945</th><th>Separatees December 1945</th></tr>
<tr><td>Total</td><td>100</td><td>100</td><td>100</td></tr>
<tr><td>Had worked full time for an employer</td><td>80</td><td>64</td><td>82</td></tr>
<tr><td>Definite plans to return</td><td>(23)</td><td>(16)</td><td>(29)</td></tr>
<tr><td>Tentative plans to return</td><td rowspan="2">(26*)</td><td>(7)</td><td>(17)</td></tr>
<tr><td>Considering returning</td><td>(11)</td><td>(13)</td></tr>
<tr><td>Not considering returning</td><td>(31)</td><td>(30)</td><td>(23)</td></tr>
<tr><td>Had not worked full time for an employer</td><td>20</td><td>36</td><td>18</td></tr>
</table>

* Wording of answer category was such as to include all those considering returning but not definitely planning to do so

Table 1 presents a comparison of the plans of white enlisted men surveyed in the world-wide cross examination of summer 1944, and those of July and December separatees at the time of discharge. Clearly, July separatees who had been employed full time just before entering the service were planning to return to their former employers in lower proportion, and December separatees were plan-

ning to return in higher proportion than were enlisted men who had been surveyed in the summer of 1944. The first question to be answered, in evaluating the use of soldiers' plans as predictors, was whether the differences reflected the effect of different characteristics within the three sample groups, or whether there had been real shifts in the plans of soldiers with similar characteristics.

An analysis of characteristics associated with plans to return to one's preinduction employer revealed that such plans were held in highest proportion among older men, those with relatively short

TABLE 2

PLANS TO RETURN TO PREVIOUS EMPLOYER, BY DURATION OF PREVIOUS EMPLOYMENT

(July and December 1945 Separatees)

(Per cent)

	LENGTH OF PREVIOUS EMPLOYMENT			
	Less than 1 year	*1–2 years*	*2–5 years*	*Five years or more*
	July Separatees			
Total	*100*	*100*	*100*	*100*
Definite plans	12	22	31	49
Tentative plans	9	11	12	16
Considering returning	20	18	17	14
Not considering returning	59	49	40	21
	December Separatees			
Total	*100*	*100*	*100*	*100*
Definite plans	18	35	47	64
Tentative plans	22	24	17	16
Considering returning	17	19	13	13
Not considering returning	43	22	23	7

service in the Army, and those longest employed in the job held just before induction. The last-mentioned variable, however, was found to account for all of the correlation between age and plans to return and most of the correlation between length of Army service and plans to return.[5] Even when groups of the same duration of employment are compared, however, December separatees consistently expressed plans to return to their previous employers in substantially higher proportion than did July separatees (Table 2). The same was found to be true when groups of identical age, duration of Army service, and education were compared.

[5] There was a considerably higher relationship between plans to return and length of Army service, duration of previous employment being held constant, among July separatees than among December separatees.

Performance of Separatees

In order to establish the relationship between plans expressed at separation and performance defined in terms of an attempt to carry out those plans, the follow-up questionnaire asked not only about current employment status but also about efforts to return to preinduction employer. The question asked and the distribution of responses of July and December separatees are reported in Table 3.

TABLE 3

RESPONSES OF JULY AND DECEMBER SEPARATEES TO QUESTION RELATING TO RETURN TO PREINDUCTION EMPLOYER AS OF TWO TO FOUR MONTHS AFTER DISCHARGE

QUESTION 6: Since you got your discharge, did you go back to work for the employer (company, person, etc.) you worked for before you entered the Army?	*July separatees (per cent)*	*December separatees (per cent)*
Total	*100.0*	*100.0*
I wasn't working for an employer before I went into the Army	22.8	12.1
Yes, I got a job with my old employer and I am working there now	19.3	38.2
Yes, I got a job with my old employer but I am *not* working there any more	3.1	4.1
I tried to get a job with my old employer, but couldn't	1.7	3.6
No, I asked about a job but decided not to work there	4.1	6.0
No, I haven't started working there yet, but I plan to go back with my old employer	3.6	4.3
No, I didn't try to go back and I don't plan to	43.0	29.1
No answer	2.4	2.6

The extent of return of July and December separatees to work for their preinduction employers was very closely in accordance with predictions made on the basis of their plans. Thus 23 per cent of the July separatees had expressed definite or tentative plans to return, and as of two to four months after discharge 22.4 per cent had returned and another 3.6 per cent still planned to return, while 1.7 per cent had tried to get their old jobs back but had been unsuccessful. In all, then, 27.7 per cent may be classified as having returned or tried to return. Again, among December separatees 46 per cent had expressed plans indicating that they were leaning toward return, and within three to four months after discharge 42.3 per cent had returned, another 4.3 per cent still planned to return,

and 3.6 per cent had tried to return but had been unable to do so. Thus, 50.2 per cent may be classified as having returned or tried to return.

Actually, the accuracy of the prediction of return to preinduction employer was probably somewhat higher than indicated by the above figures. It appears that about 2 per cent of each group of separatees was comprised of men who reported that they had returned or had tried to do so, yet who had at the time of discharge indicated that they were not employed immediately prior to their induction. In classifying plans to return, it will be recalled that men not employed full-time *immediately* prior to induction were not included, since they presumably had no real claim to jobs.[6]

It will also be noted that answers to the category "I wasn't working for an employer before I went into the service" give a considerably lower estimate of the number not gainfully employed than was given by questions in the questionnaire administered at the separation centers. (See Table 1.) Some men employed only part-time before induction answered the follow-up question in terms of their relation to their part-time employer. Some men not employed at all prior to induction checked "I didn't try to go back" instead of checking that they had not been employed; others simply did not answer the question. All of which again indicates the problem of maintaining exact specification of dimensions in asking such questions.

Prediction for the Individual Separatee

Among both July and December separatees there was a consistent relationship between plans classification and performance. The large majority of soldiers who had planned definitely to return actually did so; a smaller proportion of tentative planners returned; while a sizable minority of those considering returning and relatively few of those not considering returning actually did so.

It has been noted that December separatees planned to return (and did return) to work for former employers in much larger proportion than did July separatees. It will be noted (Tables 4 and 5) that for every category of plans with respect to returning to preinduction employer, December separatees returned in higher proportion than did July separatees. Whereas 82 per cent of the July

---

[6] An alternative method of prediction, which did not eliminate men who were not employed immediately prior to discharge, yielded estimates of 26 per cent and 47 per cent return for July and December separatees, respectively.

separatees with definite plans to return did go back to their old jobs, had attempted to go back, or were planning to do so, 85 per cent of the December separatees with definite plans to return had at least tried to carry out their plans or were planning to sustain them. Again, 11 per cent of the July separatees who originally had not planned to return to their old employers apparently changed their minds and at least tried to go back or were planning to do so, but among December separatees with comparable plans, 24 per cent returned, or attempted to do so, or were intending to return.

TABLE 4

RELATIONSHIP BETWEEN PLANS TO RETURN TO PREVIOUS EMPLOYER AND EXTENT OF RETURN WITHIN TWO TO FOUR MONTHS AFTER DISCHARGE

(July 1945 Separatees)
(Per cent)

STATUS TWO TO FOUR MONTHS AFTER DISCHARGE	PLANS EXPRESSED AT SEPARATION*			
	Definitely return	*Probably return*	*Possibly return*	*Plan not to return*
Total	*100*	*100*	*100*	*100*
Had returned to previous employer	71	44	26	7
Planning to return	9	8	5	2
Tried, could not get job back	2	3	2	2
Asked, but decided not to return	6	6	8	5
Made no effort to return	12	39	59	84
Number of cases	*296*	*131*	*206*	*524*

* By men who had worked full-time for an employer just before entering the Army

There is clear evidence that December separatees were more likely to return, regardless of their plans.

This fact would seem to reflect the differences in the labor market situation encountered by the two groups of separatees rather than marked differences in the personal characteristics of the separatees themselves. Among July separatees who at the time of discharge were somewhat undecided in their plans but were considering going back to their old jobs, a third did return to their former employers within two to four months after discharge. The other two thirds apparently found other prospects more attractive. But among December separatees in the same category as to plans, only about half found more attractive alternatives to returning. Indeed, on

the basis of performance the plans category "considering returning" for the December group came very close to being equivalent to "will probably return" for the July group.

For both July and December separatees, the cutting point for prediction of returns had represented a lucky choice, and errors of prediction on the two sides of the cutting point almost exactly balanced each other. The relation of plans to individual performance, however, indicates that there was no sharp dividing line between

TABLE 5

RELATIONSHIP BETWEEN PLANS TO RETURN TO PREVIOUS EMPLOYER AND EXTENT OF RETURN WITHIN THREE TO FOUR MONTHS AFTER DISCHARGE

(December 1945 Separatees)
(Per cent)

STATUS THREE TO FOUR MONTHS AFTER DISCHARGE	PLANS EXPRESSED AT SEPARATION*			
	Definitely return	*Probably return*	*Possibly return*	*Plan not to return*
Total	*100*	*100*	*100*	*100*
Had returned to previous employer	79	57	37	15
Planning to return	4	6	8	3
Tried, could not get job back	2	7	4	6
Asked, but decided not to return	4	10	9	7
Made no effort to return	11	20	42	69
Number of cases	*598*	*351*	*274*	*480*

* By men who had worked full-time for an employer just before entering the Army

plans categories but rather a gradation. If the war with Japan had not ended in August 1945, it is possible that considerably fewer July separatees would have returned to their former employers—and that the cutting point for prediction would have been just below the group with definite plans to return. Again, if the labor market had been any tighter for the December separatees, it is possible that more than half would have sought their old jobs and that the best estimate of total return would have included not only the men with definite or tentative plans to go back but also those who had said: "I may return but probably will *not*."

The Adequacy of Plans as a Basis for Prediction

In addition to the questions already described as the basis of classification of soldiers' plans to return to previous employers, the

"summing up" question toward the close of the questionnaire was also used as a basis for estimating the probable rate of return. The question, involving a final choice of alternatives, has already been quoted in connection with plans to return to school.[7] The first answer category was "I will probably work for the employer I was working for just before I came into the Army." This question, it will be recalled, was followed by another which asked about the degree of certainty of the soldier's plans. To be classified as having definite plans to return to work for one's previous employer, by this scheme, a man had to check the appropriate category in the question involving alternatives (Q 68) and check "I am very sure" (that this is what I will do first of all after discharge) in answer to the question relating to degree of certainty. All other men who checked that they thought they would probably return to work for their preservice employer (but who were not "very sure") were classified as having tentative plans to return. A comparison of the proportions of men classified as having plans to return to their previous employers, based upon the alternative methods of classification, is given in Table 6.

TABLE 6

COMPARISON OF ALTERNATIVE METHODS OF CLASSIFICATION OF SEPARATEES' PLANS TO RETURN TO PREVIOUS EMPLOYER

Interest in returning to previous employer	*Separatees July 1945 (per cent)*		*Separatees December 1945 (per cent)*	
First Method (direct questions)				
Definite plans to return	16	23	29	46
Tentative plans to return	7		17	
Considering returning	11		13	
Other (including men who had not been employed)	66		41	
	100		100	
Second Method (final alternatives)				
Definite plans to return	15	26	27	47
Tentative plans to return	11		20	
Other plans	74		53	
	100		100	

It will be noted that predictions based upon the alternative methods were closely similar—assuming that most men with definite or

[7] The full question is reported in Appendix A to this chapter. See questions 68 and 69.

tentative plans would carry out those plans and that roughly compensating numbers who had not planned to return would actually do so.

As would be expected, there was a high degree of consistency in classifications based on the two methods. Thus, of 1,713 December separatees who had worked for an employer just before entering the Army, the same prediction—whether return or nonreturn—was given by the two methods in 1,489 or 87 per cent of the cases. (See Table 7.)

TABLE 7

EXTENT OF CONSISTENCY OF PREDICTION USING ALTERNATIVE METHODS OF PLANS CLASSIFICATION, AND EXTENT OF RETURN TO PREVIOUS EMPLOYER FOR EACH

(December Separatees)

Prediction based on original method of classification	*Prediction based on second method of classification*	*Number of separatees*	*Per cent who actually returned**
Return	Return	841	83
Return	Nonreturn	118	49
Nonreturn	Return	106	60
Nonreturn	Nonreturn	648	26

* Or planned to return or had tried to return as of three to four months after discharge.

The two groups of men whose expression of plans was inconsistent, by the criterion of alternative classification, tended to counterbalance each other and to return to their previous employers in reasonably similar proportions.

Of the errors in prediction of return, only about a third came from the two rather small groups which appeared inconsistent in the expression of their plans. The other two thirds came from the larger groups who had consistently reported plans to return or consistently reported plans not to return. It would appear, therefore, that while manifest uncertainty of plans expressed at the time of discharge was a significant source of error in prediction, the modification of plans after discharge was probably of greater importance as a source of predictive error.

Population Characteristics and Prediction of Return

As has already been mentioned, the most important characteristic associated with plans to return to preinduction employer was the length of time a man had been with that employer. In addition to

its relationship to *plans* expressed at the time of discharge, duration of previous employment was also found to affect the extent of *actual return* by men who expressed similar plans. That is, among men with definite plans (or tentative plans, or no plans) to return to their previous employer, those longest employed went back in higher proportion than those employed for a shorter period (Table 8).

TABLE 8

RATE OF RETURN TO PREVIOUS EMPLOYER PER HUNDRED SEPARATEES REPORTING SPECIFIED PLANS TO RETURN AND SPECIFIED DURATION OF PREVIOUS EMPLOYMENT

(July and December 1945 Separatees)

<table>
<tr><th rowspan="2">DURATION OF PREVIOUS EMPLOYMENT</th><th colspan="4">RATE OF RETURN BY PLANS EXPRESSED AT SEPARATION</th></tr>
<tr><th>Definitely return</th><th>Probably return</th><th>Possibly return</th><th>Plan not to return</th></tr>
<tr><td colspan="5">July Separatees</td></tr>
<tr><td>Less than one year</td><td>70</td><td rowspan="2">50*</td><td rowspan="2">31</td><td>8</td></tr>
<tr><td>One year up to two years</td><td>79</td><td>12</td></tr>
<tr><td>Two years up to five years</td><td>79</td><td rowspan="2">61</td><td rowspan="2">41</td><td rowspan="2">14</td></tr>
<tr><td>Five years and over</td><td>89</td></tr>
<tr><td>Average rate for plans group (from Table 4)</td><td>(82)</td><td>(55)</td><td>(33)</td><td>(11)</td></tr>
<tr><td colspan="5">December Separatees</td></tr>
<tr><td>Less than one year</td><td>76</td><td>63</td><td>41</td><td>21</td></tr>
<tr><td>One year up to two years</td><td>83</td><td>72</td><td>56</td><td>23</td></tr>
<tr><td>Two years up to five years</td><td>86</td><td rowspan="2">77</td><td rowspan="2">53</td><td rowspan="2">29</td></tr>
<tr><td>Five years and over</td><td>93</td></tr>
<tr><td>Average rate for plans group (from Table 5)</td><td>(85)</td><td>(70)</td><td>(49)</td><td>(24)</td></tr>
</table>

* Combinations made where single classes contained fewer than fifty cases.

For both July and December separatees, tabulations to show extent of return, by length of Army service, by age and by education, holding constant plans and length of previous employment, revealed no significant relationships. That is, among groups of comparable plans and comparable duration of employment, older men were no more likely to return than were younger men, men in the Army less than three years were neither more nor less likely to return then were those in the Army for longer periods, high school graduates and college men were no more likely to return than were soldiers whose education had not extended beyond grade school. There was a slight tendency for those high school graduates who

had *not* planned to return to their previous employers to stand by their plans (i.e., not to return) to a somewhat greater degree than was the case among men with only grade school education, but this was the only group for which rate of return seemed to be related to education.

Type of Employer and Rate of Return

The questionnaire administered at the separation centers had asked for a report of the type of employer for whom the men had worked in their last civilian jobs—whether a private employer with fewer than fifty employees, a private employer with fifty or more employees, or a branch of Federal, state, or local government. An analysis of the extent of return to previous employers, by type of employer, of December separatees revealed that 40 per cent of the men who reported they had worked for small private employers, 61 per cent of those who reported they had worked for larger private employers, and 52 per cent of those who said they had been governmentally employed[8] actually returned to their former employer. In part, these differences reflect a difference in the plans of the three groups. Even with duration of previous employment held constant, plans to return to their preinduction employer were more frequently expressed by men who had worked for larger companies than by those who had worked for companies or individuals employing less than fifty persons. Part of the greater degree of return by the employees of larger companies, however, seems to be independent of plans expressed at separation. Consistently higher rates of return were noted for the employees of larger companies when plans and duration of previous employment were held constant.[9]

Of men who had worked for more than three years for companies employing fifty or more workers, nearly four fifths planned to return and a slightly higher proportion actually returned. Only three fifths of the men who had worked for smaller companies for a comparable time planned to return, and the same proportion actually returned. It is interesting to note that a number of industrial

[8] The self-reporting of governmental employment is known to contain a large element of error because of misconceptions held by many soldiers as to the scope of such employment. Thus many classified work in plants with war contracts as federal employment.

[9] Because of the small number of cases involved when comparing employees of large and small companies, matching length of preservice employment and plans to return, the significance of the differences between matched pairs was tested by the use of the statistical signs test after eliminating all pairs in which either member contained less than twenty cases. In six out of six pairs included, the rate of return was highest for employees of larger companies.

giants have reported the return of more than 90 per cent of their former personnel who entered the service.[10] The number of governmental employees involved was too small to permit the use of controls adequate to support any generalizations in this respect.

Occupation and Return

Inquiry into the occupational plans of soldiers surveyed in the summer of 1944 revealed that those men who had been engaged in occupations of higher prestige and skill requirements were more likely to plan to return to the same general field after the war than were those whose jobs had been at lower levels of prestige or skill. Thus 54 per cent of the skilled workers, 42 per cent of the semiskilled workers, and 32 per cent of the unskilled workers expressed plans to return to the same kind of work after the war (regardless of whether for the same or a different employer). Data on preinduction occupation were coded for July separatees but not for December separatees. For the July group it was found that skilled workers planned to return and did return to their preservice employers in slightly higher proportion than did semiskilled or unskilled workers, but no differences in plans or in performance were noted between the latter two groups. Moreover, all of the excess in rate of plans and of return on the part of the skilled workers appears to be associated with the fact that, by and large, this group had been longer employed by their preservice employer. The number of cases involved in other occupational groups was too few to permit adequate analysis of possible differences in rate of return.

The Situation Encountered after Discharge

Mention has already been made of the fact that December separatees encountered a markedly different employment situation than had July separatees. Unemployment had been very slight in July; by January 1946, it had passed the two million mark. In July less than 200,000 men were discharged; in December more than a million left the services to compete for places in the labor market. Between the two groups of separatees, the rate of entrance into jobs was considerably slower for the December group. Among those who had taken jobs, in the first month after discharge, job turnover was lower for the December group. Again, among December separatees the percentage who reported that they tried to get a job with their

[10] See, for example, the report of the Standard Oil Company of New Jersey in the pamphlet "A Generation of Industrial Peace" (1946), p. 5.

former employers but were unable to do so was considerably greater than the percentage of July separatees giving this report.

At the time of discharge, December separatees were to a considerable degree aware of the situation which they faced. The shift in their plans, as compared with those of July separatees, attests their awareness. But the data on their performance, for groups of comparable plans (Tables 4 and 5) suggest that those who found they could return, whether or not this was their primary desire, were more often interested in the security of a job than the July separatees had been. Again it must be emphasized that although the predictions made on the basis of soldiers' plans turned out to be surprisingly accurate for the two sample groups, a markedly different set of conditions after discharge might have altered greatly the actual rates of return.

SECTION II

SOLDIERS' PLANS FOR FARMING[11]

About 10 per cent of the men who enlisted or were drafted into the Armed Forces from the onset of Selective Service were employed in agriculture just prior to their enlistment or induction. But nearly another fifth of all the men in service had had at least a year's experience, full-time, at working on a farm. Many had left the farm years before the war emergency, as part of that cityward flow of man power which has gone on since the first cities were established. The city offered jobs, excitement, freedom, as contrasted with the lack of opportunity in many rural areas where the annual increment of infants was larger than the increment of job openings, where life moved at a slower, more even pace, and where the like-mindedness and community solidarity of the adult population seemed sometimes to preclude the possibility of enjoying "modern" pleasures. Other men had shifted from working on a farm only when the war emergency gave rise to the tremendous boom in factory employment and wages.

The war reduced the man power available to operate America's farms by a million and a half, but older men, women, and adolescents stepped into the breach to the extent that, with improved farm practices, the agricultural yield during the war years was greater than ever before. Moreover, it was estimated that the

[11] The main findings of the report *Soldiers' Plans for Farming after They Leave the Army* have been presented in Chapter 13 of Volume II of this series.

number of farm youth becoming available to operate or work on the farms would be more than enough to replace all of the older men dying in the next ten years. With improved technology, it was estimated that 300,000 fewer agricultural workers than were on hand in the war year 1943 could produce all the farm products which would be needed.[12]

In view of these facts, it was important to know what effect military service would have with respect to the attractiveness of farming as an occupation. Would men without farming experience develop a hankering for the quiet life of the farm where they might be free from regimentation? Would there be a tremendous demand from veterans for farm land? Several journalists painted pictures of city veterans flocking into farming—a trend which, if it materialized, could only result in disillusionment for many of those not trained in modern farming methods.

Ascertaining Plans for Farming

Most of the pretest work with respect to ascertaining soldiers' plans to farm was devoted to getting at the details of such plans—whether men planned to farm as operators, family workers, or wage hands, the general type of farming they planned to do, etc. Although some inconsistency was noted in comparing answers to several questions dealing with the general classification of plans to farm rather than to enter some other field, the problems here encountered did not seem as acute as those in fields like government employment or self-employment in nonagricultural enterprises. Perhaps the reason for this is that relatively few soldiers who did not have a background of farming experience expressed serious interest in an agricultural career.

Basic classification of plans to farm in all major surveys on this subject utilized responses to the write-in question and a direct check list contained several pages later in the questionnaire: "Do you think you will do farming when you get out of the Army?" This direct question on plans to farm was preceded by two questions on previous farming experience and followed by a battery of questions on the detailed plans of those men planning to farm. (See Appendix A, questions 48–58.)

The cross-sectional survey conducted in the summer of 1944 found approximately 10 per cent of the white enlisted men to be seriously

[12] See Carl C. Taylor, "The Veteran in Agriculture," *Annals of the American Academy of Political and Social Science*, Vol. 238 (March 1945), p. 50.

considering full-time farming as an immediate postwar job plan. Included in this 10 per cent were 7.7 per cent who expressed definite and consistent plans in answer to both the write-in and the direct check-list question,[13] 1.0 per cent who checked that they were "almost sure" they would farm yet mentioned in the write-in that they were considering an alternative plan, and 1.4 per cent whose write-in answer mentioned farming but who checked "I may do some farming, but I'm not sure." Clearly these are not scale classes, but it seemed probable that those expressing definite plans would sustain those plans in highest proportion and that those who checked "I may do some farming" would be least likely of the three groups to sustain their plans. The chief reason for this expectation was the finding of the way in which members of the three groups responded to this question: "Do you think you will make farming your lifework or do you expect to change to some other kind of work later on?" As will be seen in Table 9, nearly three fourths of the men who had expressed definite plans said they expected to make farming their lifework, as against half of those who mentioned that they were considering alternatives but thought they would probably farm, and an eighth of those who checked merely "I may do some farming."

TABLE 9

ATTITUDE TOWARD MAKING FARMING THEIR LIFEWORK AMONG MEN CONSIDERING FULL-TIME FARMING AS POSTWAR OCCUPATION

(Summer 1944)

Attitude toward farming as lifework	*Plan definitely to farm (per cent)*	*Think they will probably farm (per cent)*	*Say they may farm (per cent)*
Total	*100*	*100*	*100*
Expect to make farming lifework	73	51	13
Expect to work on farm but might change later	24	39	50
Expect to work on farm only if cannot get other job	1	6	26
No answer to question*	2	4	11
Number of respondents	*1,190*	*146*	*221*
Per cent of total sample	7.7	1.0	1.4

* Question was included in block to be answered only by men planning to farm. A high proportion of failures to answer the question is thus indicative of a lack of plans to farm.

[13] To this group were added men who failed to answer one of these questions or the other but who otherwise consistently expressed plans to farm.

The three groups also differed in the proportion having had at least a year of full-time farming previous to Army service. Among those with definite plans, 86 per cent had had at least a year of full-time experience while for the other two groups the figures were 78 and 66 per cent respectively. (Only 2 per cent of those with definite plans were wholly lacking in farming experience.)

Of particular interest in this part of the study of soldiers' plans was the specific type of farming arrangement planned. A majority of those employed on farms before induction had either worked on family farms or as wage hands, and only a third reported that they had owned or rented a farm (for cash or for a share of the crop). Among those with definite plans to farm, however, 70 per cent said they planned to operate a farm which they would own or rent. As might have been expected, moreover, those who had operated farms were much more likely to plan to return to farming than were the former wage workers on the farms.

It was found that men who said, in answer to one question, that they had in mind a farm which they planned to buy or rent, frequently indicated elsewhere that they planned to work on a family farm or for wages immediately after the war. In other words, they aspired to become farm operators but would be content for a while to work on someone else's farm. Other men indicated that if they could not rent or buy a farm they would not be interested in farm employment. Thus the estimation of the number of men who would actually farm, in a sense, required several predictions. A quotation from the report of the 1944 survey poses the problem:

"Prediction of the number of men who will actually farm cannot be made solely on the basis of the men's intentions of farming. The intentions of the men at this time may be relatively unimportant as determinants of later action unless certain conditions are met after the war. Among these conditions the following may be considered as of major importance: (1) there must not be too great a differential between opportunities open in agriculture and those offered by industry; (2) there must be farms available for those men who will go into farming only if they can be operators; (3) men counting on loans guaranteed by the government under the GI Bill of Rights in order to finance farm operation must actually be able to qualify for such loans." [14]

The problem of estimating how many soldiers might return to

[14] Information and Education Division Report No. B–131, *Soldiers' Plans for Farming after They Leave the Army*, p. 32.

farming after the war was also complicated by the fact that the use of a questionnaire which was filled out by the respondent underrepresented a segment of the Army population containing a high proportion of potential farmers—the illiterate. In surveys of plans of Negro troops, illiterates were systematically interviewed, but among white troops it was estimated that illiterates comprised only about 2 to 4 per cent and therefore they were simply excluded from consideration. In the present instance, however, it was estimated that roughly half of the illiterates were prospective farmers and farm laborers. In preparing estimates of the total number of soldiers likely to go into agricultural work after the war, allowance was made for the existence of the unrepresented illiterates, but the effect of their presence in the body of prospective farmers was unfortunately not reflected in the data on the details of plans for farming.

Postseparation Performance

Data on the proportion of separatees who had gone into agricultural work shortly after discharge were secured only for July separatees. Nine per cent of the July separatees reported that they had been farmers or farm workers just before entering the service, and a little less than 8 per cent were seriously considering returning to farming. As of two to four months after discharge, however, only

TABLE 10

RELATIONSHIP BETWEEN PLANS TO FARM AND POSTSEPARATION PERFORMANCE

(July 1945 Separatees)

(Per cent)

PLANS EXPRESSED AT TIME OF SEPARATION	STATUS TWO TO FOUR MONTHS AFTER DISCHARGE		
	Total	*Farming*	*Not farming*
Total	*100.0*	*6.5*	*93.5*
Definite plans to farm	4.7	2.8	1.9
Tentative plans to farm	2.4	1.0	1.4
Vague plans to farm	8.1	1.2	6.9
No expressed plans to farm	84.8	1.5	83.3

6.5 per cent reported that they were engaged in farming. For two reasons, it is likely that this figure does not accurately reflect the actual extent of return of July separatees to farming, however: (1) follow-up responses were secured from most men in October, when agricultural employment would have declined somewhat, and (2) some men were postponing their return to farming until the follow-

ing spring. Nevertheless, it is instructive to consider the relationship between plans and performance in so far as data are available (Table 10). Roughly three fifths of those men who had expressed definite plans to farm actually carried out their plans in the first few months after discharge, as did two fifths of the men who had expressed tentative or not wholly consistent plans to farm. Together these two groups contributed somewhat more than half of the men who were farming after discharge. Nearly all of the men who were farming had had some previous experience in this field and nearly two thirds of them said they had worked full-time on a farm for a year or more. This helps to explain the rather large group who reported in the follow-up that they were farming even though they had previously checked "No, I'm almost sure I won't farm." Most of these men were either working on their family's farm or were working for wages on someone else's farm—perhaps having taken, temporarily, at least, the first available job for which their previous experience fitted them.

In general, holding constant the basic classification of plans, it appears that men with at least a year of full-time experience at farming, men who said they wanted to make farming their lifework, and men who planned either to work on their family's farms or to operate their own farms were more likely to carry out their plans than were men with less experience, those less enthusiastic about a life career in farming, and those who had planned to work as hired hands. This latter finding is of particular interest since it had been hypothesized that those planning to operate farms might not be willing to return to agriculture if they could not manage to buy or rent a farm. Actually, many of these prospective operators were working on family farms or for wages, but they were working toward the goal which they had specified in their statements of plans.

A comparison of the extent to which either one of the classifying questions alone would have permitted a correct prediction as to whether or not the individual separatee would go into agriculture after discharge revealed that the simple mention of farming on the write-in question more often gave a correct prediction that the man would farm than did the checking of the category "Yes, I am almost sure I will farm full-time" in answer to the direct check-list question. It is possible that some of the experienced farmers who leaned toward returning to agriculture after the war, yet who were considering alternative plans, were led into greater and greater un-

certainty as they proceeded through the questionnaire. The fact that many of the men planning to farm had had less educational opportunity and hence encountered more difficulty than other men, undoubtedly explains some of the inconsistencies of responses to the questions on their plans.

SECTION III

SOLDIERS' PLANS FOR BUSINESS OWNERSHIP

Prior to any follow-up of separatees, the following hypothesis was jotted down for future evaluation: "The proportion of men who carry out their plans, or attempt to carry them out, will be highest among those planning definitely to return to their previous employers and lowest among those planning to start businesses of their own. More generally stated, the proportion of men who carry out or try to carry out their plans will be directly related to the amount of past experience and knowledge the men have in the area of their plans." This hypothesis was simply a common-sense statement, and one which was well borne out by the results of the follow-up studies. In the case of soldiers' plans for businesses and for governmental employment, plans were often based upon wishful thinking and upon misconceptions and were less highly predictive for the individual respondent than were plans in the areas already described. The reasons for this will be described in detail in the following sections.

Survey data indicated that nearly 90 per cent of the Army's male personnel entered the service from school or from jobs which they had held as employees. Only about 8 per cent of a large cross-sectional sample reported that they had been self-employed before entering the service, and nearly half of these had been engaged in farming, chiefly as tenants. Thus less than 5 per cent had been engaged in small businesses or other nonagricultural self-employment. Nevertheless, the traditional American goal of independence in a small business enterprise was strongly planted in the consciousness of many of the men in service. And when the Servicemen's Readjustment Act of 1944 offered, or seemed to offer, the prospect of a loan for getting started in business to any veteran who wanted to go into business, many servicemen began to view their dreams as plans for their postwar world. Fed-up with taking orders from others of higher rank in the service, not a few soldiers

expressed an aversion to returning to jobs where they would again be "outranked."

But dreams and intentions do not automatically lead to their own materialization. Many soldiers with a dream actually had little more than that dream as a basis for building a successful business. A large proportion, perhaps a majority, of those who were thinking about a postwar business could not be considered good enough risks to secure loans under the Federal program of aid to veterans. No one who interviewed them in order to check the realism of their thinking and planning would predict that most of these men would actually get started in businesses of their own in the immediate postdemobilization period.

One of the most troublesome problems encountered in working with statistics on small businesses relates to the matter of defining a business. Common usage is to classify as self-employed all persons whose employment and income rests upon their own enterprise and who are neither regular wage, salary, or commission workers nor unpaid family workers. In theory, at least, the self-employed may further be classified as own-account workers or as employers. All of the employers may be said to operate businesses (if professional practices are considered business enterprises) but common usage would not consider all of the own-account workers as proprietors of businesses. For example, the odd-jobs man is an own-account worker but hardly a proprietor. The picture is enormously complicated, however, by the fact that many individuals shift back and forth between own-account work, employment *of* others, and employment *by* others. As of any moment in time one can presumably classify all persons in a particular field, although if seasonal fluctuations are great, such classification may not represent the most usual status of the persons classified. In attempting to predict for the future, the problem is even greater. Painters, carpenters, truck owners, landscape gardeners, artists, and members of other occupational groups often indicated that they planned to seize whatever opportunities came for self-employment, yet might at times work for larger enterprises in their fields.

The number of employers and own-account workers in nonagricultural enterprises as enumerated in the 1940 Census of Population exceeded the estimated number of business firms (based upon data from the Census of Business, Bureau of Internal Revenue, various state governmental agencies, trade associations and other sources)

by about one million.[15] This excess is a reflection both of partnerships—two or more proprietors to the firm—and of the fact that many own-account workers did not maintain a recognizable place of business and therefore were not enumerated as firms.

In attempting to ascertain soldiers' plans for businesses, the basic criteria for categorization of a business were the mention of self-employment, a brief description of the work planned, and the soldier's checking "I plan to have a business of my own." In general, interest was here focused upon the soldier who planned self-employment which would qualify for consideration of a loan application under the provisions of the Servicemen's Readjustment Act. Own-account workers would in most instances qualify. It is apparent, then, that the fuzzy edges of the concept of "a business" added appreciably to the problem of ascertaining plans to enter into such enterprises.

Among the prospective users of research in this field there was some interest in the level of aspiration with respect to business, but for the most part, the government agencies charged with administering the loan program and with allocating the disposition of surplus property wanted to know how many veterans were likely to apply for aid and how many might be qualified to receive such aid.

Nowhere in the series of studies of soldiers' postwar plans was more difficulty encountered in presenting the data on soldiers' plans in such a way as to prevent wholly unfounded conclusions than in the reporting of plans for business ownership.

The section that follows will take up first the steps in the development of a set of questions to give a classification of plans for business ownership which would minimize the effects of suggestion and wishful thinking, and, second, the way in which the data on plans were utilized to give qualified predictions.

Ascertaining Soldiers' Plans to Own Businesses

A pretest in the spring of 1944 contained the following question and yielded the responses indicated:

[15] Compare United States Census, 1940, *Population*, Vol. 3, Introduction, page 6, and Table 78, with estimates by Howard R. Bowen in "Trends in the Business Population," *Survey of Current Business*, Vol. 24, No. 3 (March 1944), pp. 8–13. Additional problems of definition and classification in this field are discussed in another article by Bowen et al., "New and Discontinued Businesses 1940–1943," *Survey of Current Business*, Vol. 24, No. 7 (July 1944), pp. 7–13.

Do you expect that you will sometime start a business of your own?

______ I already have a business of my own	6%
______ Yes, I plan to start a business of my own as soon as I get out of the Army	16
______ Yes, I plan to start a business of my own but not as soon as I get out of the Army	40
______ No, I don't plan to start a business of my own	35
______ No answer	3
	100%

The question is obviously not adequate to get at specific plans, but it serves to indicate the high level of aspiration for proprietorship of a business among the soldiers surveyed. Interviewing revealed that many men who checked that they planned to start businesses immediately after discharge had no real plans but only hopes. Further, it was found that some men who checked "I already have a business of my own" had misinterpreted the question; they were merely reporting that they had a regular trade or line of work which they planned to return to. It also appeared that a majority of the prospective farmers were interpreting business to include farm operation. Thus the task of ascertaining serious plans for proprietorship of a nonagricultural business enterprise was clearly one beset with problems.

The use of a question presenting alternatives diminished considerably the number of men who indicated plans for a business, but even when self-employment was chosen over school and over working for an employer, there was evidence that *carefully considered plans* for business ownership were still greatly overstated. The use of a write-in question at the very start of the series of items on postwar plans seemed to be the only way of cutting down the expression of plans for a business to a group that for the most part would be similarly classified by an interview.

The major survey in the summer of 1944 used responses to the following questions to classify soldiers' plans for businesses:

28. (a) What kind of work do you think you will do *right after the war?*
Write the name of the job and describe it as fully as you can.
If you are undecided, tell what you think you *might* do.
(Four lines were left for description of plans)

32. Do you expect to work for yourself or for someone else *right after you leave the Army?* (Check one)
______ I expect to work for myself
______ I expect to work for the government
(Federal, state, city, etc.)

_______ I expect to work for a relative (father, uncle, etc.)
_______ I expect to work for some other employer (company, person, etc.)
_______ I plan to go to full-time school
_______ I plan to stay in the Army
_______ I don't have any definite plans

35. Do you plan to start a business or run a farm[16] soon after you leave the Army? (Check one)
_______ No, I don't plan to start a business or run a farm
_______ I already have a business or farm
_______ Yes, I am almost sure I will start a business or run a farm
_______ I may, but I'm not sure

Although 12.9 per cent of the men answered the direct question (Q 35) by checking "I am almost sure I will start a business" (prospective farmers having been eliminated), only 8.7 per cent also said they planned to be self-employed right after leaving the Army, and only 5.3 per cent had also indicated definite plans for a business when first asked what they planned to do. The extent to which insistence on consistency in all three questions cut down the size of the groups who indicated "plans" for business in answer to the direct question is shown in Table 11. Some of the men who had said

TABLE 11

EXTENT OF REDUCTION OF ESTIMATES OF PLANS FOR NONAGRICULTURAL OWNERSHIP* USING AS CRITERION CONSISTENCY ON A THREE-QUESTION BATTERY
(Per cent)

Successive classification	*Already own a business*	*Almost sure I will start a business*	*May start a business*	*Do not plan to start a business*
Total responses to direct question (Q 35)	*3.0*	*12.9*	*17.5*	*66.6*
Question with alternatives (Q 32):				
Expect to be self-employed	2.6	8.7	3.1	
Write-in question (Q 28a):				
Definitely plan business	(2.1)	(5.3)	(0.7)	
Considering starting business	(0.3)	(1.4)	(0.7)	
No mention of business	(0.2)	(2.0)	(1.7)	

* Men planning to farm were eliminated from the tabulation.

[16] Men planning to farm were eliminated by the use of an additional question which asked, "What kind of business or farming will you go into?"

they already owned businesses were actually planning to share management of family businesses after the war; others had been self-employed in small-scale enterprises and thought of themselves as owning "a business" even though the enterprise was not operating while they were in service and in fact its reopening after discharge was a matter of some doubt.

Taking only those men who were completely consistent and free of doubts in the expression of their plans to own businesses after discharge, it will be noted that a little over 7 per cent would thereby be classified as definitely planning a business. A crucial aspect of this system of classification was the adequacy of coding of the write-in question. Recalling that in a cross section of enlisted men roughly a third had no education beyond grade school, it will come as no surprise to learn that the responses to the write-in question were not always literate or even legible, and in a considerable proportion of cases they were incomplete or ambiguous. Perhaps the surprising thing is the care that was taken by the great majority of the men to answer the question to the best of their ability. In many instances, of course, help was provided by the Research Branch representative administering the questionnaire. Nevertheless, coding sketchy or ambiguous responses constituted a major problem both from the standpoint of accuracy and from that of time required. Where the write-in question was ambiguous—for example, when the respondent wrote merely "trucking" as the description of his postwar job plans—coding clerks were instructed to examine the entire questionnaire for clues as to the employment status planned—whether the man expected to drive a truck owned by some trucking concern or whether he planned to operate a trucking business. Among the clues, of course, were answers to the check-list items dealing with self-employment and business ownership, so that coding of the write-in was in a small number of instances not independent of the other items used in the classification of plans for a business. Although all coding was verified and a high degree of consistency of classification was obtained, the complexity of the coding operation and the inevitability of differences in interpretation and hence of classification led to a considerable amount of interviewing and experimentation with various alternative approaches to ascertaining plans.

A series of pretests was carried out in order to secure a set of questions which would permit simpler classification of plans (espe-

cially through elimination of involved coding processes) yet at the same time would minimize suggestion. The first of these pretests involved the use of three forms: one in which the soldier was asked first what he would *like* to do and then what he actually expected to do; a second in which the sequence of questions on postwar plans was introduced by the presentation of a check list of alternatives and a write-in question carried further on was designed merely as a source of data for occupational classification; and a third form in which the sequence of questions on plans was introduced by a full paragraph describing in detail the kind of information wanted in a general write-in description of plans.

It had been hoped that by first asking what the soldier would *like* to do after the war, and subsequently asking for his evaluation of the likelihood that he would actually attempt to carry out his aspiration, it might be possible to lead the respondent to a realistic consideration of his future prospects and through this to a reasoned statement of plans and expectations. Upon pretesting this form, however, it was found that many men who had thought through their plans were irritated by the implication that they might not be able to do what they would like to do, others were confused by the sequence, and still others with high aspirations sailed blithely along regardless of the questions designed to make them consider the problems to be encountered. Because of the tendency for this approach to irritate or to confuse, it was dropped from further consideration.

The other two pretest forms seemed to work well enough in so far as soldier reaction was concerned but gave somewhat different estimates of soldiers' plans. The number of cases involved was too small to establish the statistical significance of the differences, so a larger test, involving further development of the two forms, was then set up.

The primary objective of this second pretest was to establish whether a set of check-list questions which did not depend on a write-in for classification of plans would give the same estimates as a form in which the initial write-in question was of basic importance. Since the use of the write-in question was most important as a means of evaluating check-list reports of plans for a business, principal interest in the pretest was focused on classification of plans within this area.

The two forms were administered interleaved, the basic samples

comprising 1,085 cases for Form A and 1,088 for Form B. The distributions by age and education of respondents to the two forms were almost identical.

Form A first presented a check list of alternatives of employment status; Form B led off with a lengthy question asking for a full description of the soldier's plans, and followed this with a check-list question presenting alternatives of employment status. A comparison of the distribution of responses to the two similar but not identical check-list questions presenting alternatives of status suggests that responses to the question in Form B were definitely influenced by the preceding write-in question.[17] (See Table 12.)

Cross tabulation of the questions involving a choice of alternatives with the question relating directly to plans for business ownership cut down the percentage classified as planning to own a business from 25 per cent to 20 per cent using Form A and from 20 per

---

[17] The two questions are given below:

Form A

23a. You will probably want to take a vacation first of all after you get out of the Army. But after that, which one of the things below do you think you are most likely to do? (Check only one—the thing you think you are *most likely to do FIRST*, even if you should plan to change to something else afterward.)

______ Work for an employer, on salary, wages or commission
______ Go to full-time school or college
______ Farm for myself
______ Have my own business
______ Stay in the Army
______ I'm undecided what I will do

Form B

27. What type of employer do you expect to work for *right after you leave the Army?* (Check one)

______ I expect to work for a branch of the government (Federal, state, etc.)
______ I expect to work for some other employer (company, person, agency, etc.)
______ I expect to farm for myself
______ I expect to have my own business
______ I plan to stay in the Army
______ I plan to go to full-time school
______ I haven't decided what I will do

Unfortunately the differences in the wording of the two questions presenting alternatives are substantial; even though many of the response categories are almos identical it is likely that a part of the difference between the distribution of responses was occasioned by wording as well as by placement. Evidence that the write-in question led to a more careful consideration of plans is given by the fact that in answering the direct check-list question on plans for a business, respondents to Form B expressed such plans in significantly lower proportion than did respondents to Form A. It seems unlikely that this effect would have been caused by the different focuses of the two questions presenting alternatives, nor can it be accounted for in terms of differences between the two sample groups, since they were almost identical as to age, education, and preinduction status.

cent to 15 per cent using Form B. Use of the write-in question of Form B to establish complete consistency in the expression of plans for business ownership further reduced the proportion classified as having such plans to 11 per cent. Another 2 per cent of the respondents to Form B indicated such plans in answer to the check-list questions, but their write-in answers were not clear as to the employment status planned.

This pretest established rather clearly the need for starting off the sequence of questions with an item which would require some pause for consideration before giving an answer, and confirmed earlier judgments that an open-ended question would be less likely to lead the respondent into choice of an area of aspirations rather than of plans. It also demonstrated again rather clearly the difficulty of basing classification on responses to an open-ended question when such responses were frequently ambiguous because of the inarticulateness or laziness of the respondent.

TABLE 12

COMPARISON OF RESPONSES TO SIMILAR CHECK-LIST QUESTIONS PRESENTING ALTERNATIVES OF EMPLOYMENT STATUS WHEN ASKED AS INITIAL QUESTION IN SEQUENCE (FORM A) AND WHEN ASKED FOLLOWING A GENERAL WRITE-IN (FORM B)

(Per cent)

	Form A	*Form B*
Total	*100*	*100*
Work for an employer	38	47
Attend full-time school	12	11
Have own business	25	20
Farm for self (operator)	9	8
Stay in Army	3	2
Undecided	12	8
Other and No answer	1	4
Number of cases	*1,085*	*1,088*

One further check was made on the accuracy of plans classification of the questionnaire. The questionnaire was administered to small groups of men (totaling somewhat over two hundred men) who were interviewed individually after they had filled it out. The interviewer sought to arrive at a classification of the soldiers' plans by major area and by degree of certainty. Questions were raised with the soldier about apparent inconsistencies in his questionnaire responses and about some of the difficulties which might be encoun-

tered in trying to carry out his plans. In a number of instances the interviewer made his classification before having looked at the questionnaire and then compared it with the classification yielded by responses to the questionnaire. It was found that in the great majority of cases the use of the three-question battery, leading off with the write-in question, yielded identical classification with that made by the interviewer, when considering definite or tentative plans for business ownership. The major discrepancy which occurred was in the realm of vague plans, which showed up as a markedly positive response to the direct question in the questionnaire but were hardly apparent in the interview. In the interview it was easier to keep the soldier in contact with reality, and the focus of attention was on the plans that had been thought through.

Although the classification of plans yielded by the questionnaire was similar to that obtained by interviewing, the latter technique permitted a far more adequate evaluation of the realism of the plans. It was possible to check on relevant experience, on a knowledge of the financial problems involved in acquiring a business, and on the resources available to the soldier. The questionnaire used in the major survey in the summer of 1944 was deficient in getting at such information. A set of questions which would cover the major probing efforts of an interview was therefore added in order to obtain the data which seemed crucial for evaluating soldiers' plans for business ownership.

Soldiers' Plans for Business Ownership as of Summer 1944

On the basis of consistent responses to the three questions discussed above, 7 per cent of the large cross section of enlisted men surveyed in the summer of 1944 were classified as definitely planning to start a business, and another 4 per cent were classified as having tentative plans. At least a third of all enlisted men indicated some aspirations to become business proprietors. Yet even among the 7 per cent who consistently stated that they planned to start a business right after leaving the Army, there were many whose plans were not formulated in detail or were patently not based on a realistic appraisal of their own preparedness for such undertakings.

Although the report which presented the findings with respect to soldiers' plans for ownership of nonagricultural businesses explicitly stated that predictions could not be made directly from the data reported, and pointed out the prevalence of unrealistic planning,

several consumers of the research requoted the data as predictions of the number of veterans who would be likely to start businesses. As an aid to more careful interpretation of the data for planning purposes, an analysis of the problem of predicting the number of veterans likely to start businesses was prepared.[18] It focused attention on three aspects of the problem: (1) the plans of soldiers and the characteristics of those desiring to go into business after demobilization, (2) the extent of opportunities in the field of small business, and (3) the effects of changing characteristics of the population to be discharged, and changing conditions to be encountered at various phases of the demobilization process.

It was pointed out that only about 5 of the 7 per cent who had definite plans for business ownership had any relevant experience or training in the field which they planned to enter, and only 2 of the 4 per cent with tentative plans had such experience. Further, when the characteristics of men who reported plans to enter business for the first time were compared with those of men who had been self-employed in nonagricultural businesses, the former group was found to be younger and less well educated and to possess considerably less capital than the group of men who had been proprietors. On the other hand, the men planning initial business ventures were somewhat older and better educated, and possessed somewhat more savings than did enlisted men in the Army as a whole.

An analysis of recent trends in the number of small businesses revealed a net decline of about 540,000 in the early years of the war. Experts in the field of business analysis were generally optimistic that this deficit would be wiped out soon after the end of the war and that perhaps an additional half million small businesses might come into being during the five or ten years following the termination of hostilities.[19] But veterans could not count upon monopolizing the field. A quotation from the report will indicate the type of analysis undertaken:

> Competing with the veteran for business ownership will be two major groups: (1) those former proprietors of businesses who closed shop because war shortages made them unable to obtain needed stock and equipment for their continued operation or because they saw greater temporary attraction in war industry and (2) those civilians who in normal times would have turned to self-employment during the war period.

[18] Information and Education Division Report No. B–130A, *The Problem of Predicting the Number of Veterans Who Will Have Businesses of Their Own after They Leave the Army*, May 1945.

[19] See Howard R. Bowen, "Trends in the Business Population," *loc. cit.*

Relatively few urban males become employers or own-account workers before reaching the age of thirty or thirty-five. In 1940, for example, four fifths of the urban males who were self-employed were thirty-five years old or over. Among employed men under thirty, less than 5 per cent were self-employed, while among those forty-five and over, more than 20 per cent were self-employed. [1940 Census of Population: The Labor Force (Sample Statistics): Employment and Personal Characteristics, Table 12.] If one applies the age rates of self-employment indicated by the 1940 Census to the Army age distribution (in each case for urban males), one can estimate the expected increase in the total proportion self-employed associated with a given increase in average age of the population. This, of course, assumes conditions comparable to those which prevailed just before 1940. With five years aging of the Army population as constituted in summer 1944, one would expect the addition of somewhat less than 5 per cent to the total proportion self-employed—an addition of about 300,000 men.

In other words, had there been no war, and had economic conditions prevailed similar to those in the period preceding 1940, one would have expected that about 300,000 of the urban males now in the Army would have become self-employed during a five-year period. A much larger group of somewhat older urban males (men who were not drafted) would have become self-employed during the same period. And, as earlier pointed out, several hundred thousand men who had operated businesses of their own before the war and gave up such businesses to enter war industry must be counted on to return to self-employment.

If [previously existent] conditions were to prevail, then, one would expect that considerably less than half of the new business proprietors in the next five years would come from the population now in the Army. With the former proprietors now in war industry considered, one might expect veterans to number not more than a fourth to a third of the men who become self-employed in the next five years, if it were not for added motivation and financial help under the GI Bill of Rights which may induce relatively more veterans to attempt to have their own businesses.

The consensus among those who have explored the probable effects of the loan provisions of the GI Bill is that most men who can qualify for business loans under the provisions of the bill could secure such loans without this legislation. The guarantee of minimum earnings of $100 a month may be a somewhat greater aid to the prospective entrepreneur. However, relatively few enlisted men have any considerable fund of savings for equity capital; a much larger group of civilians will have available capital adequate both for equity and for operating purposes, and in general the older civilians will have a greater amount of relevant experience. It would appear, therefore, that the provisions of the GI Bill will enable some veterans to start businesses which could not have been established without such aid, but that few veterans will secure from the bill any lasting competitive advantage over nonveterans. . . .

If the above reasoning is correct, one would predict that, in the immediate post-demobilization period, the 7 per cent of the Army's male personnel who have both plans for business ownership and a fair share of the prerequisites for success might represent a maximum estimate of the number who can hope to succeed in business ventures. If many more men try to establish new businesses, a high proportion of failures must be expected.[20]

Even the final statements in the above quotation must be considered overoptimistic, however, for if applied to the total veteran

[20] B–130A, pp. 7–8.

population which would emerge from the war—roughly fifteen million—the estimated maximum of 7 per cent would yield more businesses owned by veterans than the maximum expected increase in new businesses over a ten-year period. And even though some of these veterans would be returning to businesses they already owned, or would be buying or taking over other already existing businesses, in the short run the figure seems unrealistically high. A serious fallacy was involved in temporarily losing sight of the population to which the estimate would ultimately be applied.

The analysis of the problem of predicting postwar business ownership by veterans from a knowledge of soldiers' plans concluded by noting that there would probably be a changing relationship between plans and performance at different phases of the demobilization process. Both the personal characteristics of soldiers being discharged and the economic situation encountered by the separatees might be expected to undergo progressive modification. Stress was laid upon the unfavorable conditions for starting a business as long as shortages of needed materials and supplies prevailed and jobs requiring considerably less expenditure of effort were available at attractive wages.

Business Plans of Separatees

A cross-sectional survey of troops remaining in the United States in the summer of 1945 revealed no significant change (from the 1944 survey) in the proportion planning to own businesses. Among enlisted separatees discharged from the Army in July of that year, however, less than 3 per cent of the men were definitely planning to own businesses while somewhat under 2 per cent expressed tentative plans to do so.

Early separatees were older, on the average, than men remaining in service, and in this respect they might have seemed more likely prospects to attempt business ownership. Many of the older men, however, enlisted prior to Selective Service; they were men of meager education and few job skills. They had neither the experience nor the training required for business operation, and relatively few of them planned business enterprises. Moreover, the war with Japan was continuing, and there was sufficient awareness of material shortages and the difficulties of starting new enterprises to lead some early separatees to defer their plans for proprietorship. In addition to the 4 per cent who expressed definite or tentative plans to go into business immediately after discharge, another 15 per cent

said they expected to go into business within five years after leaving the Army.

By December of 1945 there was an even greater general awareness of the problems to be faced by the prospective veteran entrepreneur, as a result of articles which appeared in service publications and in many of the popular commercial magazines. Enlisted men discharged from the Army in December expressed plans to go into business in slightly higher proportion than had the July separatees, but in considerably lower proportion than had soldiers questioned in the earlier cross-sectional surveys. There had, apparently, been a real shift in plans during the early months of demobilization.

Postseparation Performance

The follow-up of July separatees two to four months after discharge revealed that 2 per cent were self-classified as working in businesses of their own while another 2 per cent had made some arrangements to enter business and expected to get into operation within six months.[21] An additional 3 per cent reported some arrangements made but no definite prospects of entering business, and fully a third of the July separatees checked "I have been thinking of getting my own business, but have not made any arrangements yet." (See Table 13.)

These findings relating to the performance of separatees in the first few months after they left the Army were reasonably close to predictions made from the plans of the men. There was, however, considerable shifting of plans for business operation in these first few months after the men returned to civilian status. Only about two fifths of the men who had, while in the Army, expressed definite or tentative plans for a business actually made arrangements to get under way. Another fourth of the men with definite or tentative plans for a business abandoned their plans soon after discharge, while the remainder were still "thinking about owning a business" without having taken any action. These findings indicate a lower level of performance than had been predicted for those who expressed plans to own businesses.

Compensating for the men who had abandoned plans for business

[21] To classify separatees with respect to their efforts to get into business, questions were asked on (1) actual employment status at time of follow-up, (2) whether or not any arrangements had been made to buy or start a business, and (3) for those still planning businesses, the prospective date of getting into operation. For the detailed questions, see Appendix B, Questions 1, 10, 11, and 12.

ownership since they had returned to civilian life and for those who had done nothing to carry out their previously expressed intentions, were two groups who crystallized plans for business ownership after discharge. Men classified as having no immediate plans but planning business operations within five years contributed about a fifth of those who actually commenced operations or made arrangements to do so in the near future. And, men who left the Army with only vague plans and those who expressed no plans at all for business operations contributed over two fifths of all the men who were in

TABLE 13

ARRANGEMENTS FOR ENTRANCE INTO BUSINESS MADE WITHIN TWO TO FOUR MONTHS AFTER DISCHARGE BY JULY 1945 SEPARATEES, BY PLANS EXPRESSED AT SEPARATION

(Per cent)

ARRANGEMENTS AFTER DISCHARGE	PLANS FOR BUSINESS OWNERSHIP AT SEPARATION			
	Total	*Definite or tentative*	*Business later*	*Vague plans or no plans*
Total	*100*	*100*	*100*	*100*
In business	2	23	2	1
Had made some arrangements	5	17	8	3
Planned to be in operation within 6 months	(2)	(8)	(3)	(1)
Planned to enter business somewhat later	(3)	(9)	(5)	(2)
Thinking of owning business but had made no arrangements	34	31	60	29
No apparent interest in business	59	29	30	67
Number of cases	*1,650*	*65*	*241*	*1,344*

business or verging upon entry. The contribution of men who had not expressed definite or tentative plans for a business while in service was higher than had been predicted. The relationship of plans to performance for individuals is given in Table 13.

The criteria used for classification of "efforts to start operation of a business" did not include any specification of the characteristics of a "business." That is, the respondent's own self-classification, based on checking that he was "working in [his] own business" or that he had "made certain arrangements for getting a business of [his] own" was accepted regardless of the type of business enterprise described, just so long as the description seemed to reflect self-

employment. This meant, however, that some men who were simply hiring out their services on a day-to-day basis were included as having businesses of their own. Because of the difficulty of separating the own-account worker from the proprietor, and because the related benefits of the Servicemen's Readjustment Act specified "self-employment" rather than ownership of a "business enterprise," no attempt was made to eliminate the peddler and the odd-jobs man who peddled his services directly to the consumer whether to paint houses, trim hedges, or perform other similar jobs.

This fact, however, explains a significant portion of the "businesses" established by men who had not at the time of separation expressed plans for businesses of their own. As will be shown later, in discussing the characteristics of these men, the proportion of own-account workers was apparently much higher among the "non-planners" who reported themselves in business than among the group who carried out previously expressed plans for ownership of a business.

It will be recalled that the questionnaire used in the surveys of separatees contained toward its close a question asking for a final reconsideration of plans and a choice among alternatives of employment status. Use of this question, along with the succeeding item which indicated intensity of plans, yielded about the same estimate of the proportion of men planning a business as was obtained from the use of the three-question battery which included the troublesome open-ended items.[22] Moreover, this simple means of classification turned out to be as closely related to performance as was the more cumbersome method which required a major coding operation. Therefore, in ascertaining plans of December separatees for a business, only the simpler method was used.

Because other data were more urgently required in the follow-up of December separatees, the specific questions asked of July separatees about their efforts to get into business could not be included in the December follow-up questionnaire.[23] The question on employment status at the time of the follow-up was, however, available in the same form as had been used for July separatees, and hence

[22] Those classified as having definite plans for a business indicated they were "very sure" they would go into business for themselves; those classified as having tentative plans said they were "pretty sure"; all others were eliminated from consideration.

[23] The follow-up of July separatees was designed primarily to test the adequacy of plans as a basis for prediction. The follow-up of December separatees, on the other hand, was designed primarily to obtain data useful to the Veterans Administration for planning purposes.

an index of efforts to get into business was afforded by the number who had either actually gotten into operation or were engaged in attempting to do so at the time of the follow-up.

Table 14 presents a comparison of the extent to which July and December separatees managed to get into business, using responses to the question on current status as the sole criterion. It will be noted that this criterion yields a slightly higher estimate of actual business ownership by July separatees than was given by the more refined criterion used in Table 13. The explanation lies in the fact

TABLE 14

BUSINESS OWNERSHIP WITHIN TWO TO FOUR MONTHS AFTER DISCHARGE AMONG JULY AND DECEMBER SEPARATEES, BY PLANS EXPRESSED AT SEPARATION*

(Per cent)

STATUS AFTER DISCHARGE	PLANS FOR BUSINESS OWNERSHIP AT SEPARATION		
	Total	*Definite or tentative*	*Other*
	July Separatees		
Total	*100.0*	*100*	*100.0*
In business	2.8	24	2.0
Making arrangements*	0.7	6	0.5
Other	96.5	70	97.5
Number of cases	*1,650*	*68*	*1,582*
	December Separatees		
Total	*100.0*	*100*	*100.0*
In business	4.1	39	2.4
Making arrangements	1.2	4	1.0
Other	94.7	57	96.6
Number of cases	*2,098*	*95*	*2,003*

* Criteria different from those used for Table 13 See text for description

that some men who were on the verge of getting into operation (included in Table 13 in the category, "Had made some arrangements: planned to be in operation within six months") reported their current status as though they had already commenced operations.

It will be noted, despite the similar proportions planning to own a business (4.1 per cent of July separatees and 4.5 per cent of December separatees, as one may easily calculate from Table 14), December separatees had greater success in getting into operation than their fellow veterans discharged five months earlier had had. Moreover, in two respects the accuracy of prediction from plans

was greater for December than for July separatees: (1) a higher proportion of those with definite or tentative plans carried out those plans and (2) a larger proportion of those who entered business was comprised of men who had expressed plans to do so. The reason seems to lie in the characteristics of the members of the two groups of separatees.

Population Characteristics and Business Ownership

In attempting to evaluate soldiers' plans as a basis for predicting business ownership, particular consideration was given to such characteristics as age, education, relevant experience, and available capital. A comparison of the characteristics of those who achieved business ownership with the characteristics of unsuccessful planners tends to support the importance of all of these characteristics. The

TABLE 15

CHARACTERISTICS OF MEN WHO PLANNED TO OWN BUSINESSES AND OF OTHERS WHO DID ACQUIRE BUSINESSES

(December Separatees)
(Number)

CHARACTERISTICS	PLANNED OWN BUSINESSES		
	Successful in plans	*Unsuccessful in plans*	*Others who acquired businesses*
Age (total specified)	*40*	*52*	*68*
Under 30	23	36	54
30 and over	17	16	14
Education (total specified)	*41*	*54*	*70*
Grade school	4	17	19
Some high school	14	17	25
High school graduate or college	23	20	26
Relevant experience (total specified)	*37*	*48*	*
Had own business	19	12	
Other relevant preservice experience	18	28	
No relevant preservice experience	—	8	
Available savings at separation (total specified)	*33*	*53*	*63*
Under $1,000	11	28	36
$1,000 up to $2,000	8	15	18
$2,000 and over	14	10	9

* Data on experience were secured only from men who planned to own businesses

number of cases involved is small, but the data is presented because of their consistency and the magnitude of some of the differences (Table 15).

It will be noted that among men who had planned business ownership, those over thirty, high school graduates, those with previous experience acquired in their own businesses, and those with savings of two thousand dollars or more were, in the small samples available, far more successful than those younger, less well educated, less experienced, and less well supplied with savings.

An examination of the characteristics of men who had not definitely or tentatively planned businesses of their own, yet on the follow-up reported themselves as owning a business, reveals that they were in many respects closer to the unsuccessful planners than to the men who achieved success in their plans for business ownership. This fact may be regarded as further evidence that the "odd-jobs" sort of self-employment rather than proprietorship of a business establishment bulks rather large in the group. Since the separatees were not asked to describe their businesses in the December follow-up questionnaire, there is no way of eliminating the odd-jobs men masquerading as entrepreneurs. But the fact that savings of less than one thousand dollars were reported by more than half of the men who later claimed to be working in their own businesses, even though they had not previously expressed plans to enter business, would suggest that many of the business operations were on an exceedingly small scale. It is probable, of course, that a few of the men classified as successful in their business plans would fit into the same category, but the comparative characteristics suggest that this happened less frequently among the men who had planned businesses than among the others who reported that they had acquired their own businesses.

Adequacy of Plans Classification

It will be recalled that two methods of classification of plans for business ownership were used. The first utilized a battery of three questions—the lead-off write-in question, choice among alternatives of employment status, and a direct question on plans for business ownership—and was based upon consistent responses and degree of definiteness in the answers to all three questions. The second method was based upon use of a choice among alternatives of status coming at the close of the questionnaire, and an indication of the degree of certainty with which plans for business ownership were

held. These methods gave similar estimates of the number of men leaning more strongly toward business ownership than toward any other area of plans. Nevertheless, using a fairly rigorous criterion of efforts to go into business, less than half the men who had leaned most strongly toward business ownership carried out or sustained their plans in the first few months after discharge. Moreover, this group of successful planners comprised less than two fifths of the men who reported themselves to be in business or in process of making arrangements to start a business. Even if they comprised a somewhat larger proportion of the men who were proprietors of significant business enterprises, the relationship between the plans

TABLE 16

RELATIONSHIP BETWEEN ANSWERS TO DIRECT QUESTION ON PLANS FOR A BUSINESS AND POSTSEPARATION BUSINESS OWNERSHIP

(December Separatees)
(Per cent)

QUESTION: "Do you plan to buy or start a business of your own soon after you leave the Army?"	POSTSEPARATION PERFORMANCE			
	Total	*In business*	*Making arrangements*	*Other*
Total	*100.0*	*4.1*	*1.2*	*94.7*
Already own a business	2.8	1.3	0.1	1.4
Almost sure will buy or start a business	9.1	1.1	0.3	7.7
May (have own business) but not sure	22.6	0.7	0.4	21.5
Don't plan business	65.5	1.0	0.4	64.1
Total number of cases	*2,098*			

classification used and postseparation performance is considerably less impressive than in the case of plans for education or for returning to work for a preinduction employer. To what extent may this be attributed to inadequate data on and classification of soldiers' plans?

Table 16 shows the relationship between answers to the direct question on plans for a business and later performance. It will be apparent that predictions based upon this question would have greatly overstated the number of men likely to go into business in the first few months after discharge. Although more than three fourths of the men who reported themselves as working in their own businesses at the time of the follow-up had given some indication

of plans for a business in answer to the direct question, more than four fifths of those who had said they were "almost sure" they would buy or start a business failed to carry out their plans soon after discharge.

The method of classification of plans which had been used in preparing estimates of the extent of business ownership to be attempted by veterans was much closer to the mark of their early postseparation efforts than an estimate based solely on the direct question would have been. And, as has been mentioned earlier, the use of the three-question battery gave classifications which agreed in the great majority of instances with the classification derived from an interview with the respondent after he had filled out the questionnaire.

Plans and the Situation Encountered after Discharge

The basic problem involved does not seem to be the classification of plans but rather the nature of the plans or attitudes held by many soldiers. There was a great deal of interest in and yearning for becoming the proprietor of a business, and some men had fixed their eyes on this goal without any real knowledge of the problems to be encountered before the goal could be attained. Other men, equally desirous of becoming proprietors but better able to see the difficulties ahead, expressed interest in business ownership but definite or tentative plans to do something else first. If certain contingencies occurred, they would proceed in one direction; if other contingencies materialized, they would proceed in a different direction. Some men counted on a loan under the terms of the Servicemen's Readjustment Act; the denial of the loan might mean yielding their plans. Others who had not counted upon being able to finance a business apparently found that favorable circumstances existed—perhaps a chance at a partnership with another veteran—and got started sooner than they anticipated.

The follow-up questionnaire for July separatees asked the men: "If you were planning to have your own business, have you run into any difficulties you had not expected while you were in the Army?" Of the men who had planned business ownership and had succeeded in carrying out their plans, half said they had encountered unexpected difficulties. Of those men who had not expressed plans for business but who reported that they nevertheless had started businesses, only two in eighteen said they encountered unexpected difficulties. This suggests that in considerable measure these men who

had not planned businesses found themselves face to face with opportunities which crystallized their vague plans into plans of action.

To the veteran who wished to go to school, by and large, the opportunities were available. To the veteran who wished to return to his old job, by and large, the job was offered back to him. But to the veteran who wished to own a business, there were more difficult conditions to be met, arrangements to be made, material shortages to be overcome. Certainly, a period of three or four months was entirely inadequate to permit a fair evaluation of efforts toward business ownership. Yet even this brief period was sufficient to demonstrate that data on plans alone would not yield an accurate estimate of the number of men who would try to start businesses of their own.

It is worth looking into the possibility that a more accurate estimate might have been derived from some other method of estimation. Making certain assumptions, one might have projected past trends in self-employment by age groups to the veteran population in the postwar period. It will be recalled that the application of rates of nonagricultural self-employment by age, as derived from the 1940 Census of Population, suggested that five years of aging of the Army population as constituted in 1944 would, under prewar conditions, have added about 300,000 to the number self-employed. This was roughly 4 per cent of that Army population. Adding this group to the 3 per cent who had been self-employed at the start of the five-year period, one might have predicted that it would be 1949 before as many as 7 per cent of the Army's male population would be self-employed, had there been no war and no change in the general economic scene. Going on from this point, one would have had to weigh the influences of the war, likely to prevail in the postwar period, upon the self-employment tendencies of veterans. As has been indicated, this is essentially what was done.

The data on soldiers' plans were confronted with the best available evidence on probable business opportunities and on the competitive ability of veterans to capitalize on those opportunities. Although no exact figure was arrived at, this analysis did suggest that the veteran population's share in self-employment would be less than soldiers' plans indicated. Nevertheless, the data on plans were used to give a maximum estimate. In the absence of a minimum estimate, this probably resulted in more unrealistic planning, for the immediate demobilization period, in some quarters than

would have been carried out if no data on plans had been available.[24] It is still too early to say what the long-term effects of military service in World War II will be on the self-employment of the veteran population.

SECTION IV

PLANS FOR GOVERNMENTAL EMPLOYMENT

Nearly 10 per cent of the men surveyed in the large cross-sectional survey conducted during the summer of 1944 reported that they had worked in governmental jobs before entering the service. They constituted roughly one eighth of all enlisted men and officers who had been employed prior to entering the service. A considerable proportion of these men, however, were engaged in activities which were not normally governmental functions—they were employed in war plants run by the Federal government. Almost certainly some of the men who classified themselves as having been governmental employees were actually employees in private war plants holding governmental contracts. Thus, many of the men who reported themselves as having been governmental employees had held jobs which were unlikely to exist after the termination of hostilities.

Certainly the greatest attraction of a governmental job was the security that such a job might offer. Many of the men in service were old enough to have experienced the insecurity of the thirties. Since veterans received preference ratings for jobs with the Federal government and with a number of state governments, it seemed likely that many would make an effort to get jobs with one or another branch of government.

The task of estimating the number of men planning to work for governmental bodies was seriously complicated by an unexpected finding: many men had misconceptions as to what constituted a government job. Such misconceptions introduced errors both in the reporting of jobs held prior to entering the service and in the reporting of postwar plans. In the former instance, however, it was frequently possible to classify a man correctly on the basis of his job description, ignoring in such cases the type of employer

[24] On the other hand, prior to the studies conducted by the Research Branch, fantastically high estimates of veteran self-employment were reported by some prognosticators who had talked with a few men in uniform and had been impressed with the high level of aspirations for business ownership.

for whom he reported he had worked. Correction or allowance for the men's misconceptions were much more difficult when dealing with postwar plans, however. For in those instances where a man described a job which would rarely be found in governmental service, there was no way of knowing whether the particular job or the status of government employment was his chief aim. Another source of misconception made it difficult to determine whether men were planning Federal, state, or local employment: many men simply did not know what jobs were offered by what governmental bodies. Some, for example, said they wanted jobs in city governments and then specifically mentioned post-office employment. Others were under the impression that teaching jobs were offered by the Federal government.

Aside from the extent to which these misconceptions complicated the classification of the men, they were in themselves important clues to general attitudes toward governmental employment. To many soldiers, "government" apparently connoted Washington, politics, and bureaucracy. These men did not think of the varied jobs offered by the Federal government (as well as by state and local governments) in all parts of the country. By contrast, for other soldiers, governmental employment connoted the security which seems inherent in large-scale operations, whether in the postal service, ship-building, or the field of transportation and communication. It appears also that governmental regulation of certain phases of transportation and communication resulted in a fairly widespread belief that these facilities are governmentally operated. Many soldiers were simply not in a position to understand the functioning of their government, possessing as they did such meager information about it. And certainly a great number of them did not have adequate information about government jobs to enable them to evaluate the opportunities such jobs might offer.

Ascertaining Plans for Government Jobs

In addition to the troublesome problem of misconception about government jobs, the ascertaining of soldiers' plans for such jobs was complicated by the fact that here again aspirations greatly exceeded informed plans. Thus, in the major survey conducted in the summer of 1944, two fifths of all enlisted men said that they "would like" to have a government job. Included in this figure were 4 per cent who reported that they had a government job waiting for them, 12 per cent who said that they would definitely try to

get a government job, and 24 per cent who said, "I would like a government job, but I am not sure I will try to get one." Even the 16 per cent who indicated that they would have such a job waiting for them or would definitely try to get such a job was considerably greater than the number who indicated in answer to an earlier question about the type of employer they expected to work for, "I expect to work for the government"—9 per cent.

Because interest in the security presumably attaching to government jobs caused many men to be suggestible when asked if they planned to try to get such jobs, classification of plans in this field was based primarily on the pattern of responses given to three questions. Those men who indicated in answer to the general write-in question that they had definite plans for a job, who indicated when asked to choose among the alternatives of employment status or class of employer that they planned to work for the government, and who answered the direct question as to their interest in government jobs by saying that they either had such a job waiting for them or would definitely try to get a government job, were classified as having definite plans for a government job, provided that there was no inconsistency between such plans and their reported interest in returning to work for their preservice employers. The evoking of these criteria left only 4 per cent of white enlisted men classified as having definite plans for government jobs, while another 5 per cent were classified as having tentative plans for government jobs, but plans into which some note of uncertainty or inconsistency had crept.

Subsequent pretesting revealed that slight changes in the wording of the direct question relating to interest in a government job would markedly modify the distribution of responses to the question. Copies of the questionnaire used in the survey conducted in the summer of 1944, interleaved with copies of a pretest version of the final questionnaire, were administered to roughly four hundred enlisted men at a single Army post in the spring of 1945. The distribution of responses is given below along with the questions and the answer categories. Question A is the form used in the 1944 survey; question B is the form used in the final survey conducted among separatees.

A. Do you think you will try to get a government (Federal, state, city, etc.) job after the war? (Check one)

1 ____	I have a government job waiting for me which I expect to take	10%

2 _____	I am definitely going to try to get a government job	15
3 _____	I would like a government job, but I am not sure I will try to get one	25
4 _____	I would try if I couldn't get some other job I would like	15
5 _____	I am not at all interested in a government job	31
	No answer	4
		100%

B. Do you think you will try to get a job with some branch of the government (Federal, state, city, etc.) right after you leave the Army? (Check one)

1 _____	I left a government job to enter the Army, and I expect to return to it	7%
2 _____	I am definitely going to try to get a government job	10
3 _____	I will probably try to get (or return to) a government job, but I am not sure	7
4 _____	I *may* try to get a government job	30
5 _____	I am not interested in a government job	41
	No answer	5
		100%

It was felt that the term a "government job" would be more likely to evoke a stereotyped reaction than the phrase "a job with some branch of the government." Again, although many men who had been governmentally employed said colloquially, "I have a government job waiting for me," and therefore this wording had been used for the first answer category in the earlier survey, interviewing revealed that some men who checked the category meant that they had been promised a government job by some friend or relative. Thus the revision of the first answer category was in the direction of greater specificity and the elimination of ambiguity. Since the objective of these surveys was to attempt prediction of the number of men who would try to get government jobs, it seemed desirable to word the third answer category in terms of such efforts rather than in terms of general interest. In both instances the question categories afford a crude scale of interest in governmental employment. In the case of the final form of the question it was thought that the cutting points yielded by the various answer categories would be more valuable than those yielded by the former question.

In the survey of separatees, the classification of plans for governmental employment was based upon the direct question cited above and upon a question placed earlier in the questionnaire inquiring as to the type of employer for whom the man expected to work—

whether a branch of the Federal government, a branch of a state government, or a city, town, or county government. Among July separatees nearly 15 per cent said they were definitely going to seek a job in some branch of government or return to such a job previously held. But only 8 per cent had also previously indicated that they expected to work for a branch of the government.

Table 17 presents the results of a cross tabulation of the question on type of employer (Q 27) and the direct question on plans for a government job (Q 46). Most of the men who had checked that

TABLE 17

INTERRELATIONSHIP OF RESPONSES TO TWO QUESTIONS RELATING TO INTEREST IN GOVERNMENT JOBS

(July Separatees)
(Per cent)

Q 46: Do you think you will try to get a job with some branch of the government? RESPONSE CATEGORY CHECKED	Q 27: What type of employer do you expect to work for? RESPONSE CATEGORY CHECKED *Total*	*Branch of gov't*	*Private employer*	*Undecided, No employer, No answer*
Total	*100.0*	*12.1*	*49.6*	*38.3*
Plan to return to gov't job previously held	2.8	2.2	0.2	0.4
Definitely going to try to get gov't job	11.8	6.1	3.0	2.7
Will probably try	8.0	1.4	3.2	3.4
May try	36.5	1.8	20.1	14.6
Not interested in gov't job or No answer to question	40.9	0.6	23.1	17.2

they expected to work for a branch of the government gave consistent indication of interest in governmental employment when asked the latter question. It would appear, however, that the direct question exerted a strong element of suggestion on a considerable number of men who had earlier indicated either that they planned to work for a private employer or that they had no idea what type of employer they would work for. Thus responses to the earlier question indicated that about 12 per cent of the men might be expected to make a definite effort to get government jobs and that, in addition, some of the men who were undecided as to type of employer (about 20 per cent) might also become interested in government employment. Responses to the direct question alone sug-

gested that nearly 23 per cent of the men might be expected to make some effort to get a government job and that another 36 per cent *might* be interested in governmental employment. It seems apparent that a large number of the men who were not planning to go to school, to return to their previous employer, or to try to start a business—the men who would be seeking a job with a new employer—might be interested in certain government jobs but simply were not planning definitely to work for any specific employer or type of employer. Many, in fact, indicated that they wanted to "shop around" and then take the best opportunity that was open to them. One might predict that a majority of the 8 per cent who were consistent in their expression of intentions to seek or return to a government job might do so, but there was no way of estimating how many additional soldiers might decide to go after government jobs.

Because of the change in the methods of classification of plans for government jobs, it was not possible to ascertain whether or not there was a shift in plans between the summer of 1944 and the summer of 1945. On the basis of a comparison of responses to similar questions, it appears that if a change did occur it was not great. December separatees reported plans for governmental employment in roughly the same proportions as did July separatees.

Performance of July Separatees

The follow-up of July separatees contained two questions relating to efforts to get a job with some branch of the government. Because of space limitations and the more urgent need for other types of data, the follow-up of December separatees did not inquire into efforts to obtain a government job.

The basic question used to classify July separatees according to their efforts to get a government job, and the percentages of individuals who checked each answer category are given below:

Have you tried to get a job with any branch of the government (Federal, state, city, etc.) since you got out of the Army?

______	Yes, and I am already working for a branch of the government	8.6%
______	Yes, I have my application in for a government job and will take the job if they offer it to me	7.1
______	I did apply for a government job, but am not sure I want it now	4.1
______	No, I haven't tried yet, but I plan to	29.1
______	No, I am not interested in a government job	45.4
	No answer	5.7
		100.0%

Roughly a fifth of the July separatees applied for government jobs in the first two to four months after discharge. Since many Federal agencies were building up staffs to handle phases of veteran aid during demobilization, there was a large supply of temporary Federal jobs available to veterans discharged early in demobilization, even though total Federal employment was soon to decline rather sharply. Notwithstanding this fact, the postal service far outstripped other governmental employers of July separatees. Next in importance were jobs connected with Army and Navy installations, including separation centers. City and county maintenance and protective jobs were the chief source of public employment other than Federal.

Prediction for the Individual Separatee

Among men who had consistently indicated at the time of separation that they intended to seek a government job, three fifths had applied for a government job and another fifth said they still planned to apply when surveyed in the follow-up of July separatees. A fifth had apparently lost interest within the first two to four months after discharge. Among those who had indicated an intention to seek government employment in answer to one of the two classificatory questions but had not given a consistent answer to the other question, a fourth applied for government jobs within the first two to four months after discharge and over two fifths reported to the follow-up that they had not "tried yet, but [still] plan to" try to get a job with some branch of the government. Even among the large group who had stated at separation either that they were not interested in a government job or merely that they might try to get a government job, fully a seventh of the men did apply for some form of public employment within the first few months after discharge. Table 18 gives the relationship between plans for governmental employment and postseparation performance.

Because of the magnitude of the segment of the population reporting at separation, "I *may* try to get a government job" (more than a third of the total survey group), the men classified as having only vague plans or no plans for a government job actually contributed a little more than half of the applicants for governmental jobs soon after discharge. The men with tentative or inconsistent plans (16 per cent of the survey group) contributed a fifth of the applicants for government jobs, while the 8 per cent who had expressed definite and consistent plans contributed slightly over a fourth of the applicants.

Thus, predictions that a given individual would or would not seek a government job yielded a low degree of accuracy when based upon the soldier's plans. Moreover, there is no reason to believe that much better predictions *could* have been made for individuals at the time of Army discharge. An interviewer, ascertaining what specific expectations the individual separatee had, with respect both to job wanted and to government employment opportunities available, might have scored a slightly higher proportion of successful predictions, but it is doubtful that any considerable gain in accuracy could have been made. It would appear that a firsthand knowledge

TABLE 18

POSTSEPARATION EFFORTS TO OBTAIN A GOVERNMENT JOB BY PLANS EXPRESSED AT SEPARATION

(July Separatees)
(Per cent)

POSTSEPARATION EFFORTS	PLANS FOR GOV'T JOB AT SEPARATION		
	Definite and consistent	*Tentative or inconsistent*	*Vague plans or no plans*
Total	*100*	*100*	*100*
Employed in gov't job	30	9	6
Applied for gov't job	29	16	8
Would take job if offered	(17)	(9)	(5)
Application rejected	(2)	(1)	—
Not sure still want job	(10)	(6)	(3)
Plan to apply	20	44	27
Not interested in gov't job	15	26	54
No information	6	5	5
Number of cases	*138*	*273*	*1,236*

of available opportunities—both in private industry and in government—which was secured from active job-seeking after discharge, was the crucial element in leading to a final decision between government and private employment. The man who knew he wanted government employment could more often be counted upon to make an active search for such employment than could the man simply interested in "a job"; but, even for the more interested men, there was frequently the problem of applying and qualifying well before the job became available. Many of these men indicated in their free comments at the close of the follow-up questionnaire that they simply could not wait for months while Civil Service registers were being prepared. On the other hand, there were many men

sufficiently interested to apply for a government job even though they were temporarily returning to work for their previous employers, or for some other private employer, permitting the possibility of making a shift later on.

The prediction of the proportion of all veterans who might seek government jobs has, nevertheless, been reasonably accurate, according to statistics on Federal employment. As of August 1948, veterans and persons entitled to veteran preference numbered 846,001 out of 1,895,443 Federal employees in the continental United States.[25] Over 560,000 of these were male veterans of World War II.

It will be recalled that 4 per cent of the soldiers surveyed in 1944 expressed definite plans and another 5 per cent expressed tentative plans to secure a government job; of these, 60 per cent were primarily interested in Federal employment. In the two years following virtually complete demobilization of the wartime Armed Forces, roughly 4 per cent of all male veterans secured Federal employment. Thus, the high level of interest in Federal jobs expressed by soldiers and separatees seems definitely to have carried through, despite the fact that jobs in private industry were plentiful.

SECTION V

GENERAL OBSERVATIONS AND CONCLUSIONS

The detailed discussion of the problems of attempting to ascertain soldiers' plans and of the relationship of expressed plans to actual postseparation performance has demonstrated the complexity of one type of prediction problem. The particular prediction problem dealt with here may not soon be encountered again. But this *type* of problem will probably become increasingly frequent. Large-scale social planning seems more and more inevitable, and large-scale planning, in a democracy, involves predicting how people will react to a given program. Efficient operation without coercion means anticipating needs before they become acute, and providing means of meeting them. Every time a program is proposed which will affect the opportunities or alternatives available to people who are faced with a problematic situation, planning will require an assessment of the probable choices which people will make in the changed situation.

[25] According to a press release made by the Civil Service Commission on October 8, 1948.

What generalizations can be drawn from the present study as a basis for better predictions in other studies of the same type? It may be noted that generalizing for such purposes is itself a predictive operation. One predicts that if a given prediction problem is analyzed in a stated way, using specified techniques, the resulting predictions will be more accurate than if a different type of analysis had been used. The problem takes on added interest after the failure of the public opinion polls to predict the outcome of the 1948 Presidential election. It seems safe to assume that the polls erred significantly not only in sampling but also (and perhaps primarily) in gauging the disposition of the voter. One may ask, then, what techniques are known that would permit more accurate prediction? No attempt will be made here, however, to analyze the errors of the polls or to apply the generalizations here formulated to the problem of predicting elections.

It should, nevertheless, be noted that when these generalizations are applied, they become statements of prediction, and they are subject to errors of prediction. They may be expected to give accurate results to the extent that the factors involved in the new problem area are similar to those which were encountered in the prediction of soldiers' postseparation behavior. Many more studies of this type will be required in order to elaborate and qualify properly the generalizations derived from the present study.

The main methodological conclusions which may be drawn from the studies of soldiers' postwar plans may be divided into two parts: (1) those relating to the technique of ascertaining plans and assessing intentions; and (2) those relating to the prediction problem itself.

Soldiers' Plans and Questionnaire Methodology

It has already been observed that questionnaires rather than interviews were used to obtain the basic data, not because the questionnaire method seemed intrinsically better, but because it was the only feasible method of securing adequate data from a large cross section of the Army. Without protracted intensive interviewing, a suitable questionnaire could not have been formulated. Interviewing revealed the extent of uncertainty and some of the elements considered by the soldier in formulating his plans. The basic nature of these plans was that they were complex, shifting orientations to a future situation which could only be imagined. They were not unitary fixed attitudes which could be "measured"; they were not orientations which could be given by an automatic

response to a stimulus question asking what the soldier planned to do. They involved reasoning through and selecting from a number of alternatives—and if they were to be realistic they had to involve an awareness on the part of the soldier of what he would want at some future date when the environmental pressures upon him would be quite different from those experienced while in service.

It seemed imperative, therefore, that the questionnaire be designed in such a way as to encourage the respondent to reflect on various alternatives and to state his general expectations before seeking answers about detailed aspects of his plans. In addition, a knowledge of how he reacted to questions seeking the details of his plans was itself essential in order to classify those plans as to realism and consistency. In this sense the questionnaire was designed similarly to the so-called "funnel-shaped interview" used in nondirective interviewing. In the questionnaire those items dealing with details of plans to enter a given field after the war were arranged in blocks to be answered only by men who were seriously contemplating such entry.

The effects of context as relating to general questionnaire organization may be briefly summed up. Whether one led off with a battery of questions on previous experience, on personal characteristics, or on postwar plans, did not apparently have any marked effect upon the reporting of plans, other things being equal.[26] But the way in which the batteries of questions on postwar plans were developed did affect the reporting of plans. An attempt to work from aspirations to reasoned statements of intentions was confusing and failed to bridge the gap between aspiration and realistic possibilities. A relatively nondirective question used as a lead-off in the area of plans was more successful in minimizing the reporting of aspirations than were other types of questions. Questions relating to the degree of certainty of plans as reported gave a basis for analyzing factors associated with definiteness of intentions.

Working from the general to the specific and detailed, the questionnaire was designed to obtain considerable detail on plans to enter fields of high prestige value. In this way data on plans could

---

[26] It may be noted, however, that in the pretest which utilized four general questionnaire designs administered to matched groups of soldiers, answers to a series of questions on personal adjustment were affected by placement in the questionnaire. When carried at the close of the questionnaire, which dealt primarily with postwar plans, the questions were answered much more negatively than when placed early in the form. A possible explanation is that the consideration of postwar plans was depressing to some men.

be evaluated as a basis for prediction in the light of the experience, training, and financial resources possessed by the respondent. Moreover, after having led the respondent to weigh these factors in his own thinking, these blocks of detailed questions gave him an opportunity to state whether he might do something else, immediately after separation, before attaining the status aspired to. Thus, he was given an "out" which did not damage his prestige but suggested a temporary compromise. This proved to be an extremely effective way of screening out a number of men who had otherwise consistently expressed plans for business ownership and farm management. In the first few months after separation, very few of these men did enter the activities they aspired to.

In addition to considering the effects of question order and overall questionnaire design, it may be worthwhile to summarize the relative advantages and disadvantages of the different types of questions asked in these studies.

Advantages and Disadvantages of the Various Question Forms

The open-ended, nondirective question. The nondirective question does not force the respondent to answer in a narrowly specified dimension. All questions which require the respondent to choose among a set of suggested responses are excluded, as are all other questions which suggest the phraseology of response. In a sense, as will be discussed shortly, a question which does not indicate the broad field of subject matter in which the response is sought is a meaningless question for classification purposes (except as it may be valuable for diagnostic purposes in psychological study). There was, however, a great range in the degree of specificity of the questions used in attempting to ascertain soldiers' plans.

The most general question of this sort—i.e., most completely nondirective—was attempted in an early pretest. The question, which was aimed to obtain a classification of plans by employment status and broad occupation was: "What do you think you will do right after the war?" This question proved to be of little value because nearly half of the respondents answered in such terms as "settle down to a quiet life," "raise a family," etc. In an interview such irrelevant answers (from the point of view of classification of employment plans) would of course be followed up by a directive probe such as, "I mean your plans for work or school—what are they?" In the questionnaire, at least this minimum specification of dimensions was necessary in the original question.

The nondirective question has the great advantage of not suggesting an answer to the respondent and (ideally) of getting at the respondent's own frame of reference. It has several disadvantages when used for a questionnaire to be filled out by the respondent, particularly if the respondent be a person with only grade school education and no great facility for setting down words on paper. The question calling for a written answer must be coded in order to classify the response. This is a time-consuming process and one which may introduce a considerable element of judgment and hence a considerable danger of errors of classification. Moreover, if the question is nondirective, some respondents will not respond along the desired dimension. And if dimension is made sufficiently explicit, the element of suggestion may occur.

Two forms of the general open-ended question were used in the major versions of the questionnaire.

> A. What kind of work do you think you will do right after the war? (Write the name of the job and describe it as fully as you can.)
>
> B. What kind of work do you plan to do after you get your discharge and take a vacation? In your own words describe what you think you will do for the first few years of your civilian life. Tell whether or not you have some definite job lined up, write the name of the job, and describe it as fully as you can. If you do not have a definite job lined up, tell what kind of work you will probably do. If you plan to take one job at first and change jobs later, tell what you plan to do first and also what you will try to change to. If you are undecided, tell what you have been thinking you *might* do.
>
> (Try to write fully enough about your job plans for the next few years so that we will understand them about as well as you yourself understand them at the present time.)

It will be noted that possible plans for school attendance were not mentioned in either version of the write-in because it was found that the mention of school served to increase the inconsistency between expressed plans for school and those for employment. Nevertheless, the exclusion of any mention of school does serve to bias the question slightly against the expression of educational plans, so for this reason the write-in was not used in the analysis of plans for school attendance.

The earlier version of the write-in (A) sought to require as little reading as possible, in order not to discourage men with little education. Subsequent pretesting showed, however, that the second form (B) would elicit a fuller response which lent itself more readily to unambiguous coding as to employment status and occupation.

Even in the case of the fuller answers, 2 per cent were ambiguous as to employment status and 13 per cent were ambiguous as to broad occupational category (single digit code), excluding from consideration an additional 13 per cent considering alternative plans. In neither case was it possible to use the write-in as a basis for ascertaining for a majority of respondents whether or not they were planning to return to a previous employer.

The direct check-list question. The direct check-list question specifies the precise dimension along which response is wanted and, except for nonresponse, permits an unequivocal classification of all members of the population with respect to responses along this dimension. "Do you think you will *actually* go to full-time school right after you leave the Army?" gets a direct self-prediction with respect to full-time school. There is little danger that the respondent will answer in terms of his job plans after he completes school (although ambiguities may of course occur in the case of poorly worded questions or areas of ignorance). Further, the question is simple and nonresponse is held to a minimum. There is no coding problem, except for the man who writes a qualification of his answer, and these are few.

But the direct check-list question does have disadvantages, not the least of which is the large element of suggestion it contains for some men, particularly if such a question is the first one asked in a sequence of questions on plans. A series of direct questions relating to plans for school, for a business, and for returning to work for preinduction employer showed a substantial number of men who were inconsistent, saying first that they would go to school and then that they would return to work for their former employer or do something else which would conflict with plans for full-time school. Such inconsistency could be resolved only by getting a direct indication of choice between the various alternatives under consideration.

The direct question might either offer a "yes" or "no" choice, present different aspects of a problem (e.g., contingencies), or attempt to get at the intensity of plans.[27] The following three variations of basically the same question were used in pretests to develop a battery of questions on educational plans:

A. Regardless of what you want to do, do you think you will actually go back to full-time school or college after the war?

_____ Yes
_____ No
_____ Undecided

[27] Contingency questions will be discussed in detail in a later paragraph, as they involve a special problem.

B. Regardless of what you would *like* to do, do you think you will actually go back to school or college after the war?

_____ Yes, to *full-time* school or college
_____ Yes, to *part-time* school or college
_____ No, I don't think I will go back to school or college
_____ Undecided

C. Do you expect to go back to full-time school or college after you leave the Army?

_____ Yes, I am almost sure to go to full-time school or college
_____ I may go to full-time school or college but I'm not sure
_____ I am planning to go to part-time school
_____ No, I don't think I will go back to school or college

Question A appeared to be somewhat frustrating to the man interested in attending part-time school or college after discharge. In some instances it caused such men to check "undecided" rather than "no," and in a few instances caused men who elsewhere indicated plans for part-time school to check "yes," although they were clearly not planning to attend full-time school.

Question B avoids the frustrating element of question A. On the other hand, the interpretation of the "undecided" category becomes dubious unless one is using other questions to indicate whether the indecision is between full-time school and part-time school or perhaps between part-time school and no school at all. As one item in a scale relating to educational plans this item proved satisfactory.

Question C avoids both the frustrating element of question A and the interpretation problem of question B. But the introduction of the second category relating to plans for full-time school still leaves a problem of deciding whether individuals who check this category are more interested in attending school than in some other possible course of action after discharge.

Choice among a check list of alternatives. The check-list question presenting alternatives was of greatest value in providing the criterion of classification for individuals considering alternative courses of action but leaning more heavily to one than to any other. It had certain additional qualities to recommend it: namely, it could be used to call to mind alternatives which were under consideration by the respondent but which may not have seemed to him to be included under the specification of the question, and it could be used to spot respondents who were basically undecided as to employment status, without having to do an elaborate coding job. On the other hand, it did not wholly avoid the suggestion factor since there appears to have been a tendency to check prestige items rather than items under primary consideration in more sober expectations.

This was especially true in the instance of pretest forms where the choice of alternatives preceded a nondirective question calling for a write-in answer.

Although the choice among alternatives seems to have provided the most satisfactory means of establishing the general plan toward which the respondent was leaning most strongly at the time of the survey, it has been noted that this question did not necessarily provide the best cutting point for purposes of prediction. A man leaning strongly toward an unrealistic plan but considering also a realistic plan—judging realism in terms of prerequisites of background and experience—was in many instances more likely to shift his orientation in the direction of greater realism after discharge. Unfortunately, the number of cases involved for any specific pair of alternatives was too few to establish actual probabilities in these studies.

The position in the questionnaire of a question presenting a choice among alternatives was found to affect the distribution of responses, but the differences in predictive value of the two sets of responses were not consistent. It had been thought that the choice among alternatives when presented late in the questionnaire would entail more careful thought and would therefore be more highly predictive than when the choice was presented early in the questionnaire. This hypothesis received support from the higher predictive value of the terminal question in the case of plans for business ownership, but the reverse was true in the case of plans for school. The value of the terminal question was also attested, however, by the fact that it yielded as good a prediction of returns to work for a previous employer as did the two direct questions on this subject. Perhaps most important, it was found that by the use of a final question involving a choice of alternatives, a reasonably good classification of plans to enter areas of high prestige value could be made at an enormous saving of effort as compared with the coding of write-in questions and the cross tabulation of a large number of questions relating to single areas of activity. While the data do not support a clear-cut generalization as to the predictive superiority of such a question when carried late in the questionnaire, the possible advantages seem to outweigh the disadvantages. On the positive side—in support of the notion that more accurate predictions may be anticipated from responses to the final question—is the fact that the respondent has been given a chance to review the crucial factors in the criterion situation and to weigh his own capabilities, limitations, competitive position, etc. On the negative side, there is always a possibility

that the confronting of the respondent by certain conditions or contingencies which he has not previously considered will tend to distort his immediate decision. That is, he may react by overemphasizing the importance of certain positive or negative factors when he first considers such factors, although after careful thought he might decide they were much less important to him.

Questions Presenting a Statement of Specific Future Contingencies

Whether or not the matter was made explicit in the questions asked, the plans expressed by the soldiers were to a considerable degree contingent plans. "*If* my old employer is still in business—" "*If* I am able to get a large enough loan—" "*If*—"

Theoretically, it would have been possible to present some of the most important contingencies in a series of questions and to evaluate plans under each for all soldiers. This was actually done in the first survey which attempted to ascertain plans for school, as has been discussed in detail, and in subsequent efforts to ascertain interest in staying in the Army after the war. It would have been extremely cumbersome, however, to present such a series of contingencies for every area of interest for prediction, and, moreover, it is doubtful whether a majority of the men would have been able to escape a feeling of confusion if forced to evaluate their plans under each separate contingency.

But another more serious problem underlies the use of contingency questions. There is evidence that, by and large, responses to such questions do not provide a basis for predicting how the respondent would act, given the materialization of the contingency, but rather reflect merely whether or not the respondent considers the specified conditions as more or less favorable than those otherwise assumed to exist. In fact, the statement of conditions which seem at first glance very favorable may trick some individuals into making a prediction that they would utilize the favorable circumstances, whereas in reality other factors make such behavior exceedingly improbable. For example, when asked, "Do you think you will actually go back to full-time school or college after the war?" 7 per cent of the men questioned in the summer of 1943 said "yes." Most of these men also replied that they would go back whether or not the government aided them. Another 18 per cent said they would go to full-time school if the government provided aid. The government did make available such aid, but the

proportion thereafter stating that they would return to school was increased by very little. Many of the additional 18 per cent who had thought they would go back if this favorable contingency materialized apparently found that in reality government aid was not the deciding factor. One might, in fact, say that in the face of the stress on the desirability of increased education, some men had rationalized a previously formulated decision not to go back to school in terms of not being able to afford going back without government aid, whereas the real reasons for not going back lay elsewhere.

Attitude Scales and Prediction

At any given instant in time, one might assume the theoretical possibility of ranking soldiers in a series of finely graded classes as to the degree of their readiness to follow some specific course of action at some future date (the assumption of the existence of an attitude scale). In the case of the present research, the best possible data on soldiers' intentions as an aid in predicting entrance into any particular field of endeavor after the war would then seem to be given by their ranking on such a scale. Other things being equal, the higher the ranking, the greater the probability of actually carrying out the stated intention.

To some extent, it *was* possible to rank individuals in a *small* number of classes along an assumed continuum, either by the use of a single question whose answer categories furnished degrees of intensity (e.g., question on plans to return to work for former employer), by the use of one question involving degree of intensity and one involving statement of choice (e.g., final choice of alternatives, followed by question on degree of certainty of plans), or by the use of a battery of questions found by scale analysis to approximate the distribution of responses obtained when only a single dimension was involved (e.g., original questions on plans to attend school). Such rankings were admittedly crude, but in every instance where they were attempted, with one exception,[28] it was found that the higher the scale category, the greater was the proportion who carried out their plans. This finding may be taken as evidence that whether or not attitudes may be "measured," rough assessments of stated intentions to act along some specified dimension may possess sufficient validity to permit graded predictions.

---

[28] This was in the case of July separatees' school plans, where the small sample of "tentative" planners returned in slightly higher proportion than "definite" planners. Almost certainly this was simply a sampling fluctuation, since the finding was reversed for December separatees.

Nevertheless, the assumption that individuals may, at least in theory, be ranked in an indefinite number of classes as to the degree of their readiness to follow a specific course of action ought to be scrutinized. More specifically, let us consider the attempt to rank soldiers as to their intentions to attend school. Let us say that the objective was to devise a scale containing a minimum of ten or twelve categories, each one representing a greater degree of intention or determination to attend school than the category next below it. The first step was to find out how a large number of soldiers thought and felt about the possibility of attending school after the war. What were the attractions of attending school, what were the conditions that would make them more or less likely to go, what assessments were made of the situation likely to prevail during demobilization, what were the major uncertainties? Having discovered that soldiers' thinking was concerned with the availability of job opportunities on the one hand or government aid for education on the other hand, probable age at discharge, the possibility of marriage, and a number of other factors, the researcher attempted to write questions which would reflect the influence of these contingencies on the decisions to be made. It has been pointed out that the answers to contingency questions cannot be taken at face value; but what of the possibility that such answers could provide a basis for ranking individuals along a single dimension—intention or determination to attend school? If a person says he would go to school under the most unfavorable circumstances (relative to school attendance) surely he should plan to go under more favorable circumstances. In general, this was found to be true, but, as has been observed, individuals differed as to whether a particular circumstance was considered to be favorable or unfavorable. Men who had started college were less deterred from planning to return by the thought of a long war than were those who had merely finished high school; men of independent means who were considering attending college as one of several alternatives were less influenced by the availability of government aid than were the less well-to-do men whose considerations involved college as an alternative. In so far as questions involved contingencies or values which had a differential appeal to different men, the questions would not permit the ranking of individuals along a single continuum. Where such rankings were approximated through question batteries using contingency questions, the approximation was made either by combining response categories which destroyed the meaningful compo-

nent of the contingency question—e.g., where individuals who said they would go to school in the face of an unfavorable situation were combined with those who said they would go only if the unfavorable situation did not materialize—or where the differential effect of contingencies or values was relatively slight. Thus a question presenting a contingency having a markedly different effect on ranking than other questions would have to be dropped out of use if one wanted to use a battery of questions as a scale.

There are, of course, other types of questions which may be asked as part of a battery to yield an expression of plans with respect to any area. There are, as has been observed, various ways of putting the question of plans for school without mentioning contingencies, and in general these various ways yield somewhat different classifications. It has been observed that a direct check-list question yielded a larger proportion reporting school plans than did a choice among alternatives. By and large the responses to these questions fulfilled the requirements of scale theory that all individuals positive on the more extreme statement (choice among alternatives) should likewise be positive on the less extreme. Through the use of such questions, which do not call to mind different values attaching to the object or different situations to be encountered, it would again be theoretically possible to arrive at a type of ranking which seems unequivocal for a majority of individuals. Moreover, the correlation of any outside variable with the scale scores derived from such a ranking is precisely the same as the multiple correlation of that outside variable with the items in the scale.[29]

Realistically, however, one cannot ignore the fact that the decisions to be made will be made on the basis of weighing values and contingencies. Admitting that scale scores will give the best possible predictions to be made from a group of items which form a scale, it is precisely the items which do not fit into a scale which pose the crucial problem for classification. It is only by assessing as best one can the probable materialization of contingencies and the implications of the responses to questions getting at such contingencies and values that meaningful prediction for the individual can be made.

It may be worth while to present the pros and cons of the various types of questions in tabular form. Again, it must be emphasized that these are tentative generalizations; one cannot predict that they will hold true in other studies where a different set of variables

[29] See Chapter 1, this volume.

is involved. Nevertheless, they may be worth consideration when similar studies are undertaken. (See Table 19.)

In so far as the predictive value of responses to a single question is concerned, it is not possible to state categorically that one type of question is always superior to another. Where plans with respect to a given dimension can be reasonably well evaluated for nearly all members of a sample through the responses to a general write-in question, it would appear that such responses are in general more highly predictive than are responses to a single direct check-list question, especially for areas where aspirations run high and clearly

TABLE 19

CONSIDERATIONS RELATING TO THE USE OF VARIOUS TYPES OF QUESTIONS IN SELF-ADMINISTERED QUESTIONNAIRES FOR PREDICTING SUBSEQUENT BEHAVIOR

Type of question	*Advantages*	*Disadvantages*
Semi-nondirective (open-ended)	Ideally, gets respondent's own frame of reference Permits qualification and elaboration Avoids suggestion Avoids prestige element	May evoke irrelevant response dimensions Does not permit ranking of all respondents along any single dimension Gets little data from poorly educated or inarticulate persons Difficult to code
Directive, single dimension (check list) a. No stated contingency	Permits unequivocal classification of nearly all respondents for specified dimension Few coding or editing problems	May impose arbitrary dimensions and obscure meaningful ones May overemphasize area under consideration Favors prestige items Does not indicate first preference if alternatives are also under consideration
b. Contingency (additional considerations)	Permits rough assessment of importance of specified contingency to plans	May overstress importance of specified contingency
Choice among check list of alternatives	Provides criterion for unequivocal classification of individuals considering alternatives Permits explicit review of alternatives Provides convenient means of estimating the number undecided as to future status	May lead to overstressing of prestige items Cutting points may not be best for prediction
Scale battery of check-list questions	Permits rough ranking of individuals with respect to given continuum	Ignores differential effects of contingencies Does not provide cutting points

formulated plans are relatively rare. A general write-in question will permit sample-wide evaluation of plans with respect to a given dimension only when that dimension is meaningful in the planning of nearly all of the individuals sampled. Thus, in the case of such dimensions as "school—nonschool," "farm—nonfarm," and "self-employed—working as employee" which actually seem to have been dichotomized in soldiers' thinking (and in so far as a reasonably complete response to the write-in was given), it was possible to make a fairly adequate classification of plans from the write-in alone. But where the dimension was one which was of secondary importance in the thinking of a large segment of the survey group—as in the case of government jobs—or one which involved the combination of several very strong components—as in the case of plans to return to a previous employer, involving as it did job rights, occupation, personal ties, etc.—it appears that plans could best be assessed by the use of direct questions which clearly specified the dimensions along which classification of plans was desired. Responses to such questions required supplementation as well as a knowledge of the more meaningful dimensions, but at least they permitted very rough gauging of potential interest. Where plans simply were not crystallized with respect to an area, no attempted assessment of plans could have provided an adequate basis for prediction.

Observations of the Prediction Problem

The objectives of these studies of soldiers' plans were such that several different kinds of data were required. For certain types of program planning there was a need for the best possible estimate of the number of veterans who would actually enter specified fields—full-time school, agriculture, etc. At the same time there was a need for estimates of the amount of pressure which would be put upon existing facilities, along with a knowledge of the characteristics of men planning to enter fields which were sure to be overcrowded. Thus there was a need on the one hand for an assessment of the potential opportunities available to veterans and on the other hand for a study of the various other factors which might affect the actual activities engaged in by veterans subsequent to discharge.

Although predictions for the individual were not required as such, the absence of any preexistent trends which might be extrapolated meant that the estimates sought had to be built up from an analysis of the problems of decision faced by individual veterans, and of the

relationship between expressed plans and other personal characteristics and performance in the situation to be encountered after discharge. If plans were to provide a basis for reasonably accurate predictions for the individual, it was noted that at least four basic conditions would have to be met: (1) that the large majority of soldiers be at least roughly classifiable as to their plans; (2) that the method used to assess plans yield both reliable and valid classifications, so that the estimate would have some stability and would reflect a real tendency to carry out the stated course of action; (3) that both the individual and the situation(s) upon which his plans were predicated would not change so markedly as to lead to a major shift in plans after assessment; and (4) that the individual's plans (including his definition of the situation) be sufficiently realistic so that the plans could be maintained when the actual situation materialized.

In so far as these conditions were not met, errors of individual prediction tended to arise, but such errors, particularly those resulting from a failure to attain the first two conditions, did not necessarily invalidate estimates of the proportions of soldiers who would follow specified courses of action, since to some degree the errors were compensating. As had been anticipated, some men with definite plans yielded those plans in the face of unexpected difficulties, while others with only tentative or vague plans decided to follow the course deserted by others.

Statements of Intention as Predictive Items

Even if a complete and accurate knowledge of the individual's plans and intentions could be secured by questionnaire, the predictive value of such plans and intentions would remain in question. Dollard[30] has made a tentative analysis of the conditions under which opinions and statements of intention permit accurate prediction of behavior. Certain of the concepts he uses will be helpful here. He uses the term "origin situation" to refer to the situation or process in which the opinion or intention was originally formulated. In the case of most soldiers who reported definite plans when surveyed, it may be assumed that the origin situation preceded the actual questioning. A decision had been made—perhaps years

[30] John Dollard, "Under What Conditions Do Opinions Predict Behavior?" *Public Opinion Quarterly*, Vol. 12, No. 4 (Winter 1948–49), pp. 623–632. This paper was originally read before a joint meeting of the Washington Statistical Society and the Washington chapter of the Institute of Mathematical Statistics, March 9, 1944.

before, perhaps only a few weeks or even days before—as to what they would try to do after discharge. When the "test situation"—the questionnaire—was encountered, this decision was reported. For other soldiers, the test situation was encountered before definite plans had been crystallized, and the questionnaire may actually have been the basis for reaching a crystallization of plans. About a sixth of all soldiers surveyed were unable to reach even a tentative formulation of plans at the time of discharge, although they reported their various interests and proclivities with respect to a number of possible activities. For all of these groups of men there lay ahead a third important situation—the one around which prediction was centered. Dollard has called this the "criterion dilemma." For the soldiers surveyed, the criterion dilemma involved the problem of settling back into civilian life as students, as employees, or as self-employed workers.

Dollard has attempted to specify certain conditions favorable to accurate prediction from statements of opinion or intention. They may be divided into two types: those relating to the characteristics of the individual and those relating to the congruence of the three situations in which opinions or intentions are formulated, expressed, and put to the test of action. Individual characteristics favorable to self-prediction are: (1) freedom from conflicts of which the individual is unaware, (2) verbal skill for the analysis of one's own inclinations and intentions, and (3) habitual action on the basis of previous thought. Situational conditions favorable to prediction are: (1) an origin situation closely approximating the criterion dilemma, (2) a test situation uncomplicated by extraneous rewards or punishments, (3) a test situation closely approximating the criterion dilemma in presenting the relevant factors eventually to be considered, and (4) no new origin dilemma—i.e., no shift in intentions—subsequent to the test situation. With the exception of the last item (which is essentially the same as the third of four basic conditions laid down for successful prediction in this study), the above conditions all relate to the (fourth) general requirement previously stated, that plans and intentions must be "realistic" if they are to permit accurate prediction. On the basis of the studies of soldiers' postwar plans, it is possible to supplement certain of these conditions enumerated by Dollard.

First, considerable support is given to Dollard's common-sense statement that "a man can best predict what he will do in a future situation if he has been in about the same situation before." Obvi-

ously, few soldiers had previously faced the problem of deciding what they would do when discharged from the service, but those who expressed plans to return to an activity they had already experienced were more likely actually to do so than were those who planned to try a different activity.. Thus men who had been in school and planned to return realized their plans in higher proportion than did those who before the war had been employed but planned to shift to school after the war. Again, plans being comparable, those men who had been self-employed were more likely to resume that status than were men who had worked for an employer. The same tendency for experience to be associated with more accurate self-prediction was noted in the case of soldiers returning to work for a previous employer, in the case of soldiers returning to farming, and in the case of those seeking government jobs.

In all the above instances, one may say that the man who had had previous experience was able to make a more realistic judgment as to what he would do, not merely because he had a truer knowledge of the criterion situation but also because there were continuing pressures upon him to utilize the benefits of his previous experience. This is reflected in the fact that even when men did *not* plan to resume the activity in which they had been engaged, they tended to do so in larger proportion than did other men with comparable plans.

Dollard observes, as separate conditions, that prediction is favored if "no new origin dilemma intervenes between the test and criterion situation" changing the individual's intentions, and "if the test question explicitly presents the conflict, i.e., anticipations of rewards and punishments, of the criterion dilemma." In a sense, both of these conditions relate to the previous condition that the situation postulated in formulating plans be similar to the situation subsequently encountered. The end of the war, subsequent to the statement of plans but prior to the follow-up of July separatees, posed a new origin situation, calling for the reformulation of job plans, at least by some men, because it changed job opportunities in the criterion situation. Likewise, marriage on the part of the man who had planned to return to school might produce, or be part of, another type of origin dilemma. The test question could approximate the criterion dilemma only if it brought to mind the possibilities of the GI Bill and the difficulties posed by marriage for the potential student. Yet, if these possibilities and difficulties had not been

recognized in the process of formulating plans, originally, then the test question might itself pose a new origin dilemma.

It is at this point that one may question or at least desire to qualify Dollard's formulation. To say that "prediction is favored providing that no new origin dilemma intervenes between the test and criterion situation" is in a sense merely to say that people will do what they said they will do if they don't change their minds first. But it carries also the implication that once the individual has stated his intentions, he remains in this respect static unless some crucial experience leads him to change his intentions. Actually, where true dilemmas are involved, there is usually a good deal of vacillation in the individual. In any given test situation, the individual is subject to moods and the influence of recent events. Anyone who has faced a major decision relating to a change in career, where the pros and cons very nearly balanced, knows that a decision made tentatively and not yet regarded as a commitment may be subject to considerable modification and fluctuation. Although the present study does not adequately document this point, it is the writer's belief that a large chance factor is involved in both the test and the criterion situation. Much depends upon the mood and recent experiences of the subject at the times when he is faced with reporting his plans and with making the actual decision.

In a sense then, the concept of origin situation is an artificial one in cases where a definite determination to act in a given way has not been formulated. In such cases, one attempts to assess a prevailing orientation, not to present a stimulus question which will evoke a ready-made response.

To restate Dollard, prediction of choice or decision at some future date (or of the carrying through of choices already formulated) is favored if the respondent is led to consider all of the alternatives which will be open to him and the consequences involved in the case of each alternative. It is probable that almost any prediction which contains some degree of ego-involvement is tinged somewhat by the aspirations of the predictor.[31] If the respondent can be confronted with some of the problems to be encountered in the criterion dilemma, the effects of such wishful thinking may to some extent be minimized.

[31] Early in the war, it was found, for example, that men most desirous of avoiding overseas or combat service tended to predict a considerably shorter war than did men who felt that they were ready for combat. See also Ogburn's finding on the prediction of football scores and college grades: W. F. Ogburn, "Studies in Prediction and the Distortion of Reality," *Social Forces*, Vol. 13, No. 2 (December 1934), pp. 224–229.

Wishful thinking must be considered not only in connection with the aspiration level of the respondent but also in connection with the implicit assumptions he makes as to the development of contingencies. It has been stressed at several points that widespread unemployment of veterans would probably have produced far greater errors in prediction than were actually observed. Yet widespread unemployment actually was anticipated by a great many veterans. Toward the close of the war, concern about future jobs led the list of personal anxieties reported. If, then, soldiers anticipated unemployment, why should the self-predictions which they made materialize most accurately in a situation of high employment and relative prosperity? One answer seems to be that optimism or wishful thinking led many soldiers who expected a postwar depression to discount the effects to themselves. Whereas 79 per cent thought *most soldiers* would find it "very hard" or "fairly hard" to get the kinds of jobs they wanted after the war, only 46 per cent thought they *personally* would find it "very hard" or "fairly hard" to get the kinds of jobs they wanted. Thus, between optimism with respect to their own competitive positions and general aspirations that led to reaching beyond immediate realities if poor economic conditions should prevail, many soldiers who predicted poor conditions were actually planning for prosperity.

The experience with contingency questions, however, casts some doubt as to whether prediction is *always* favored "if the test question explicitly presents the conflict, i.e., anticipation of rewards and punishments, of the criterion dilemma." As already suggested, rewards or punishments considered in the test situation may lead to a different decision from what would be arrived at when those factors were faced in the total context of the criterion situation. It is patently impossible to sketch out all of the relevant factors relating to an occupational choice, in questioning respondents as to their plans. In mentioning some factors, to the exclusion of others, it is conceivable that one will elicit responses even more divergent from those which would be made in the criterion situation than if the respondent were left to consider the factors regarded as relevant by him.

Other things being equal, a person can better predict his choice with respect to an activity when it is to be primary than when it is to be secondary. In a more limited sense, choice of a full-time activity can be more accurately predicted than can choice of the same general course if it is to be pursued as a part-time activity. The

prediction of enrollment of full-time students was considerably more accurate than was that of part-time students. Again, prediction of the amount of postseparation vacation to be taken was considerably less accurate than was prediction of the primary activity pursued. Choices which involve a primary activity must, to some extent at least, be squared with choice of place of residence, available opportunities, and possible alternative activities; but a secondary or part-time activity can be chosen without ruling out other activities in which one is interested. Thus, less rigorous thinking is required in arriving at the decision in the instance of secondary activities. Moreover, it is probable that aspirations or pressures to "give the right answer" are more likely to affect the reporting of a future action which need not be brought into congruence or possible conflict with other actions.

Where prediction from intentions opposed the prediction which would have been made on the basis of personal characteristics and background factors alone, accuracy of prediction was lower than when intentions and characteristics were both favorable to performance. But plans or intentions definitely held (positively or negatively) permitted more accurate predictions for most subgroups of the population studied than could have been made on the basis of a knowledge of personal characteristics and background factors alone.

Support is given to the common-sense generalization that a person is best able to predict his subsequent behavior if he feels fairly certain of his plans. Clear evidence was afforded in every area of prediction here studied, that men relatively sure of their plans were far more likely to carry out those plans than were those who reported some uncertainty.

There were no consistent indications that better educated men predicted their activities more successfully than men with less formal schooling, among individuals holding definite plans. On the other hand, the men with relatively little formal schooling were more likely to be undecided in their plans, so that somewhat poorer predictions were made for this group as a whole, in comparison with high school graduates.

Finally, the present study has clearly indicated that an analysis of the future situation for which one is attempting to predict behavior is a crucial element in the prediction process. Knowledge of the relationship between intentions or attitudes and behavior in one situation does not insure accurate actuarial predictions even for seemingly similar situations. One might say that accurate predic-

tion for the individual is favored to the extent that he is able to manipulate or control the pressures upon his behavior in the criterion situation. Where opportunities to carry out plans and intentions are good, a much higher proportion of correct predictions will be made than where unforeseen pressures demand compromises and reevaluations.

If the primary interest in these studies had been the prediction of individual behavior *as such*, one would certainly have wanted information on the specific situations to which the men were returning—on the views of close friends and relatives who might influence the returned veteran, on the specific opportunities and disabilities to be encountered, etc. But in this study, and probably in most studies of this sort where attitude data are utilized for planning a program, generalized assessments, both of intentions and of situational factors, have to suffice as a basis for qualified predictions.

Appendixes to Chapter 16

Appendix A. Questionnaire Used to Survey Soldiers at Separation Centers

1. What is your Army rank or grade? (*Put a check on the line in front of your answer like this* √)
   1 ______ Private or Pfc
   2 ______ Corporal or T/5
   3 ______ Sergeant, any grade

2. How old were you on your last birthday? (*Please fill in the blank below*)
   __________ years old on my last birthday

3. Are you: (*Check one*)
   1 ______ Single
   2 ______ Married
   3 ______ Widowed, divorced or separated

4. How many children do you have? (*Check one*)
   ______ I'm not married
   ______ Have no children
   ______ Have one child
   ______ Have two children
   ______ Have three or more children

5. How far did you go in school? (*Check only one answer—the highest grade of school you completed.*)
   ______ Less than 6th grade
   ______ Finished 6th grade
   ______ Finished 7th grade
   ______ Finished 8th grade
   ______ Some high school but did not finish
   ______ Graduated from high school
   ______ Some college but did not finish
   ______ Graduated from college

6. How many years was it between the time you left *full-time* school and the time you came into the Army? (*Check one*)

   ______ I came into the Army less than 3 months after I left full-time school
   ______ 3 months up to 1 year
   ______ Over 1 year up to 2 years
   ______ Over 2 years up to 3 years
   ______ Over 3 years up to 5 years
   ______ Over 5 years

7. What is your race? (*Check one*)

   1 ______ White
   2 ______ Negro
   3 ______ Other

8. How long have you been in the Army? (*Check one*)

   1 ______ 6 months or less
   2 ______ Over 6 months up to 1 year
   3 ______ Over 1 year up to 2 years
   4 ______ Over 2 years up to 3 years
   5 ______ Over 3 years up to 4 years
   6 ______ Over 4 years up to 5 years
   7 ______ Over 5 years up to 10 years
   8 ______ Over 10 years

9. (a) Have you served outside the Continental United States in this war? (*Check one*)

   ______ Yes
   1 ______ No

   (b) Have you been in actual combat in this war? (*Check one*)

   2 ______ No
   3 ______ I have been under enemy fire, but not in actual combat
   4 ______ I have been in actual combat

10. What branch of service were you last in before you came to the Separation Center? (*Check one*)

    1 ______ Engineers
    2 ______ Field Artillery
    3 ______ Quartermaster Corps
    4 ______ Infantry
    5 ______ Medical Department
    6 ______ Armored Force
    7 ______ Signal Corps
    8 ______ Coast Artillery
    9 ______ Anti-Aircraft
    11 ______ Air Corps
    12 ______ Transportation Corps
    — ______ Ordnance Department
    — ______ Chemical Warfare Service
    — ______ Military Police
    — ______ Tank Destroyer
    — ______ Other. *Which one?* ____________________

11. In what State of the United States did you spend most of the last five years of your civilian life before you came into the Army?

    ____

    (*Write the name of the State on the line above*)

12. Where were you living most of the last five years of your civilian life before you came into the Army? (*Check one*)

1 ______ On a farm
2 ______ In the country but not on a farm
3 ______ In a village with fewer than 2,500 people
4 ______ In a small city with between 2,500 and 25,000 people
5 ______ In a city with between 25,000 and 100,000 people
6 ______ In a large city with more than 100,000 people

13. What living arrangements did you have before you came into the Army? (*Check one*)

1 ______ Lived in the home of my parents (*or wife's parents*)
2 ______ Lived in the home of other relatives or friends
3 ______ Lived in a boarding or rooming house
4 ______ Rented an apartment
5 ______ Rented a house
6 ______ Owned a house
7 ______ Other. *What?* ______________________

14. Which of the following were you doing *just before you came into the Army?* (*Check as many as you were doing just before you came into the Army.*)

1 ______ Working full-time for an employer, for salary, wages or commission
2 ______ Going to full-time school or college
3 ______ Farming (*either with your family or for yourself*)
4 ______ Had my own business (*did* not *receive salary, wages or commission*)
5 ______ Looking for a job
6 ______ Working part-time for an employer
7 ______ Going to part-time school or college
8 ______ Something else. *What?* ______________________

15. In the LAST job you had before coming into the Army, were you working for a company or person or branch of the government? (*Check one*)

1 ______ A company or person with fewer than fifty employees
2 ______ A company or person with fifty or more employees
3 ______ A branch of the Federal Government
4 ______ A branch of a state government
5 ______ A city, town or county government
6 ______ Other employer. *What?* ______________________
7 ______ I was *not* working for an employer

16. How long did you work for your last employer before you came into the Army? (*Check one*)

1 ______ Less than six months
2 ______ Six months up to a year
3 ______ One year up to two years
4 ______ Two years up to three years
5 ______ Three years up to five years
6 ______ Five years or more
7 ______ I was not working for an employer

17. *On the average* how much money were you earning in your LAST job? (*Fill in* one *of the blanks below.*)

Weekly pay . . . . $______________ per week

or

Monthly pay . . . . . . $______________ per month

______ I did not receive any regular pay

18. (a) What kind of work were you doing just before you came into the Army? What was your job called, what business or industry did you work in, and *what did you do?* (*For example: "Sales clerk, waited on customers in a grocery store" or "Weaver, operated a loom in a cotton textile mill."*)

____
____
____
____

(b) About how long did you do that *kind of work*, either on your LAST job or other jobs?

About ________ years and ________ months

THIS IS THE MOST IMPORTANT QUESTION IN THE QUESTIONNAIRE.
READ IT *THROUGH* VERY CAREFULLY.

19. What kind of work do you plan to do after you get your discharge and take a vacation? In your own words describe what you think you will do for the *first* few years of your civilian life. Tell whether or not you have some definite job lined up, write the name of the job, and describe it as fully as you can. If you do not have a definite job lined up, tell what kind of work you will probably do. If you plan to take one job at first and change jobs later, tell what you plan to do first and also what you will try to change to. If you are undecided, tell what you have been thinking you *might* do.

Try to write fully enough about your job plans for the next few years so that we will understand them about as well as you yourself understand them at the present time.

____
____
____
____

20. How sure do you feel that after leaving the Army you will actually do the kind of work you described just above? (*Check one*)

1 ______ I'm almost sure I will
2 ______ I probably will but I may do something else
3 ______ I'm really not at all sure what I will do

21. Before you came into the Army, did you ever do the kind of work you think you will *most probably* do after you leave the Army? (*Check one*)

1 ______ No, I never did that kind of work
2 ______ No, but I studied to do that kind of work before I came into the Army
3 ______ No, but I learned that kind of work in the Army
4 ______ Yes, I did that kind of work in civilian life
How long did you do that kind of work?
For about ________ years and ________ months

22. Regardless of whether you have definite plans or not, which one of these things do you think you will most probably do FIRST—right after you leave the Army? (*Check one—if you plan to do more than one, check the one you expect to spend the most time at*)

1 ______ I will probably work for an employer, for wages, salary or commission
2 ______ I expect to work on a relative's farm (*father's, uncle's, etc.*)
3 ______ I expect to work in a relative's business (*father's, uncle's, etc.*)
4 ______ I plan to go to full-time school or college
5 ______ I plan to run my own business or farm

23. Do you plan to work for an employer the FIRST thing after you leave the Army and have a vacation? (*Check one*)
   1 ______ Yes, I plan to work *full-time* for an employer
   2 ______ I plan to work *part-time* for an employer. *What else will you do?*
   ____
   3 ______ I don't have definite plans, but I will probably work for an employer
   4 ______ No, I don't plan to work for an employer

24. Do you think you could get work with the employer you worked for before you came into the Army, if you wanted to? (*Check one*)
   1 ______ Yes, I'm almost sure I could
   2 ______ Yes, I think so, but I'm not sure
   3 ______ No, I probably couldn't
   4 ______ I'm sure I could not
   5 ______ I was not working for an employer before I came into the Army

25. Do you think you *actually will* go back to work for the same employer you worked for before you came into the Army? (*Check one*)
   1 ______ Yes, I'm almost sure I will
   2 ______ Yes, I probably will
   3 ______ I may, but I probably will *not*
   4 ______ No, I'm quite sure I won't
   5 ______ I was not working for an employer before I came into the Army

26. If you do *not* plan to go back to work for your last civilian employer, do you know of some other job you can get and plan to take? (*Check one*)
   1 ______ I *do* plan to go back to my last civilian employer
   2 ______ Yes, I know of a job I can get and I plan to take it
   3 ______ I know of a job that I can probably get and I *may* take it
   4 ______ No, I shall have to look for a job
   5 ______ I don't expect to work for an employer right after I leave the Army

27. What type of employer do you expect to work for in your FIRST job after you leave the Army—a company or person or branch of the government? (*Check one*)
   1 ______ A company or person with fewer than fifty employees
   2 ______ A company or person with fifty or more employees
   3 ______ A branch of the Federal government
   4 ______ A branch of a state government
   5 ______ A city, town or county government
   6 ______ Other employer. *What?* ______________________________
   7 ______ I have no idea what type of employer I'll work for
   8 ______ I'm not planning to work for an employer right after I leave the Army

28. As far as you can tell now, do you plan to stay in the first job you take for at least five years, or do you expect to change to some other job or type of work within five years after you leave the Army? (*Check one*)
   1 ______ I plan to keep my first job for at least five years, if I can
   2 ______ I plan to change to something else within five years after I leave the Army
   3 ______ I don't have any plans to change, but I might want to

29. If you plan to leave your first job within five years, what do you plan to do after you leave your first job?
   ____
   ____
   ____

30. Do you plan to buy or start a business of your own soon after you leave the Army? (*Check one*)

1 ______ I already own a business which I expect to operate after I leave the Army
2 ______ No, I don't plan to buy or start a business
3 ______ Yes, I am almost sure I will buy or start a business
4 ______ I may, but I'm not sure

(*If you already have, or plan to have your own business, answer the questions below* [*questions 31–37*]. *If not, go on to question 38.*)

---

31. What kind of business do you plan to operate and what do you yourself expect to do in this business? ____

32. If you have had some experience in this kind of business, where did you get your experience? (*Check one*)

1 ______ In my own business
2 ______ In my family's business
3 ______ In a business owned by someone else
4 ______ In the Army
5 ______ I have not yet had any experience in such a business

33. Do you expect to start a new business or go into a business that is already running? (*Check one*)

1 ______ I plan to start a new business
2 ______ I plan to buy a business or a share in a business that is already running
3 ______ I plan to take over a business that now belongs to someone else in my family
4 ______ I already own a business in my own name

34. Do you expect to own the business by yourself or do you expect to go into partnership with someone else? (*Check one*)

1 ______ I expect to own the business by myself—in my own name
2 ______ I expect to go into business as part-owner with my father or some other relative
3 ______ I expect to go into partnership with someone not related to me

35. What do you estimate that the total cost (both cash and credit) of getting under way will be? (*The total amount for buying or renting a building, tools, supplies, or whatever you will need, but not counting your living expenses while you are getting started*)

About $____________________
______ I haven't yet made any estimate of what the business would cost

36. Have you made any actual arrangements yet for getting and operating a business, or is your idea still in the planning stage? (*Check one*)

1 ______ I already own a business
2 ______ I have been thinking about running a business but have not made any actual arrangements yet
3 ______ I have already made certain arrangements for getting and operating a business. *What arrangements have you made?* ____________________

37. Do you plan to go into business for yourself as soon as you leave the Army or will you do something else first? (*Check one*)

1 ______ I plan to work for someone else for a while first
2 ______ I plan to go to school for a while first
3 ______ I plan to go into business for myself as soon as I leave the Army
4 ______ Something else. *What?* ______

(*Everyone should answer these questions*)

38. Do you feel that you have had as much school or college as you want, or would you like to go back to school or college after the war? (*Check one*)

1 ______ I have had as much school or college as I want
2 ______ I would like to go back to *full-time* school or college
3 ______ I would like to go back to *part-time* school or college

39. Regardless of what you would like to do, do you think you will actually go back to school or college after you leave the Army? (*Check one*)

1 ______ Yes, to *full-time* school or college
2 ______ Yes, to *part-time* school or college
3 ______ No, I don't think I will go back to school or college
4 ______ Undecided

(*If you do plan to go to school or college either full-time or part-time, answer the questions below* [*questions 40–45*]. *If not, go on to question 46.*)

---

40. What kind of school are you planning to attend? (*Check one—if you plan to attend more than one kind of school, check the kind of school you will go to* first.)

1 ______ High school, general academic
2 ______ High school, vocational
3 ______ Trade school (*Diesel, electrical, etc.*)
4 ______ Business school
5 ______ College or university
6 ______ Other. *What kind?* ______

41. Just what course or type of course do you plan to take? (*Give the name of the course and a brief description of what it covers*)

______

42. What school or college do you plan to go to?

______

______ I haven't decided yet

43. Do you plan to enroll in school or college when the next school term begins? (*Check one*)

1 ______ Yes, I plan to enroll when the next school term begins (*or even sooner*)
2 ______ No, I don't plan to enroll until later

When *do* you plan to start school?
Month ______ Year ______

44. How long do you expect to go to school or college? (*Check one*)

1 ______ Less than six months
2 ______ Six months up to one year
3 ______ One year up to two years
4 ______ Two years up to three years
5 ______ Three years up to four years
6 ______ Four years or more

45. What kind of work do you plan to do after you finish school?

____

____

(Everyone should answer these questions)

46. Do you think you will try to get a job with some branch of the government (*Federal, state, city, etc.*) right after you leave the Army? (*Check one*)

1 ______ I left a government job to enter the Army, and I expect to return to it
2 ______ I am definitely going to try to get a government job
3 ______ I will probably try to get (or return to) a government job, but I am not sure
4 ______ I *may* try to get a government job
5 ______ I am not interested in a government job

47. If you plan to get a government job, what will you do if you have to wait a while after you apply before you actually get the job (either until you take a Civil Service test or for some other reason)? (*Check one*)

1 ______ I don't plan to apply for a government job
2 ______ I plan to do something else until I can get a government job. *What?*

____

____

3 ______ I will just wait until I can get a government job
4 ______ I am almost sure I will be able to begin work in a government job right away
5 ______ I don't want a government job if I have to wait to start work

48. Have you ever worked on a farm for as much as a year? (*Check one*)

1 ______ Yes, full-time for a year or more
2 ______ Not full-time, but I grew up on a farm
3 ______ I worked on a farm, but less than one year
4 ______ No, I never worked on a farm

49. If you worked on a farm, whose land did you farm on the last time you worked on a farm? (*Check one*)

1 ______ I never worked on a farm
2 ______ I owned the land in my own name
3 ______ My family *owned* the land and I was farming with them
4 ______ My family *rented* the land and I was farming with them
5 ______ I rented the land from my family, but was farming entirely for myself
6 ______ I *rented* the land from some other landlord
7 ______ I was working on someone else's farm for salary or daily wages

50. Do you think you will do farming when you get out of the Army? (*Check one*)

1 ______ Yes, I am almost sure I will farm full-time
2 ______ I plan to farm part-time—that is, to have a steady job off the farm at the same time. *What job?* ______________________

____

3 ______ I may do farming full-time, but I'm not sure
4 ______ I may do farming part-time, but I'm not sure
5 ______ No, I'm almost sure I won't farm

(If you plan to farm, either full-time or part-time, answer the questions below [*questions 51–58*]. *If not, go on to question 59.)*

51. Do you think you will make farming your life work, or do you think you may want to do some other kind of work later on? (*Check one*)

1 ______ I expect to make farming my life work

2 ______ I expect to work on a farm for a while, but I might later change to some other sort of job
3 ______ I expect to work on a farm only if I cannot get another sort of job I want
4 ______ I expect to farm part-time only

52. Where do you plan to get the land to farm on *right after you leave* the Army? (*Check one*)
1 ______ I *already own* the land in my own name
2 ______ I *plan to buy* the land
3 ______ My family *owns* the land and I will farm with them
4 ______ My family *rents* the land and I will farm with them
5 ______ I will *rent* the land from my family but will farm entirely for myself
6 ______ I will *rent* the land from some other landlord
7 ______ I will work on someone's farm for salary or daily wages
8 ______ I'm not sure yet whose land I will farm on

53. Where do you expect to get the equipment and power (tractor or horses) to farm with after the war? (*Check one*)
1 ______ I already own the equipment in my own name
2 ______ I expect to buy the equipment
3 ______ My family owns the equipment and I will farm with them
4 ______ My family rents the equipment and I will farm with them
5 ______ I will rent the equipment from my family, but will farm entirely for myself
6 ______ I will rent the equipment from someone else
7 ______ I will work on someone's farm for salary or daily wages
8 ______ I am not sure yet whose equipment I will farm with

54. If you expect to farm after the war with land and equipment rented from someone else, what kind of rent will you pay? (*Check one*)
1 ______ I don't expect to rent land or equipment from anyone
2 ______ I will pay rent in cash
3 ______ I will pay rent with a share of the crops
4 ______ Other. *What?* ______________________________
5 ______ I am not sure what kind of rent I will pay

55. If you plan to rent or buy land in your own name *right after you get out of the Army*, do you know what land you will farm on? (*Check one*)
1 ______ No, I have no idea what land I'll farm on
2 ______ I know about where I'll farm but don't know just what farm
3 ______ I think I know the land I'll farm on
4 ______ I know exactly what land I'll farm on
5 ______ I *don't* plan to rent or buy land

56. Will you farm by yourself right after you leave the Army, or will you do something else first? (*Check one*)
1 ______ I plan to do something other than farming for a while first
2 ______ I plan to farm with my family first of all
3 ______ I plan to work on someone else's farm for wages first of all
4 ______ I plan to rent or buy a farm right after I leave the Army

57. If you plan to buy land or equipment, about how much money do you expect to pay for it? (*Total cost, not just cash payment*)
About $ ____________________
______ I don't plan to buy land or equipment
______ I haven't yet made any estimate of what it would cost

58. What equipment do you plan to buy right after you leave the Army?

____

______ I don't plan to buy equipment

(*Answer all questions from here on.*)

59. At the present time, how much money do you have saved or invested (in cash, bonds, etc.) that you can use to help get started back in civilian life after your discharge? *Do* not *count your mustering-out pay.* (*Check one*)
 1 ______ Less than $100
 2 ______ $100 up to $200
 3 ______ $200 up to $500
 4 ______ $500 up to $1,000
 5 ______ $1,000 up to $1,500
 6 ______ $1,500 up to $2,000
 7 ______ $2,000 up to $5,000
 8 ______ $5,000 or more

60. How clear an idea do you have of what is in the "GI Bill of Rights"? (*Check one*)
 1 ______ Very clear idea
 2 ______ Fairly clear idea
 3 ______ Not such a clear idea
 4 ______ Not at all clear about it

61. Do you think you will try to borrow money under the provisions of the GI Bill of Rights after you get out of the Army? (*Check one*)
 1 ______ Yes, I am quite sure I will try to
 2 ______ I may, but I'm not sure
 3 ______ No, I don't think so
 4 ______ I don't know enough about the GI Bill to say

62. If you think you may borrow money under the provisions of the GI Bill, for what purpose would you want to borrow it?

____

____

______ I don't plan to borrow under the GI Bill

63. About how much do you plan to borrow? (*Check one*)
 About $ __________
 ______ I'm not sure how much I will borrow
 ______ I don't plan to borrow

64. Do you now own a home in your own name? (*Check one*)
 1 ______ Yes
 2 ______ No

65. What living arrangements do you expect to have *right after you leave the Army?* (*Check one*)
 1 ______ I expect to live in the home of my parents (or my wife's parents)
 2 ______ I expect to live in the home of other relatives or friends
 3 ______ I expect to live in a rooming or boarding house
 4 ______ I expect to rent an apartment
 5 ______ I expect to rent a house
 6 ______ I expect to live in my own home
 7 ______ Other. *What?* ______________________________

____

66. After you have gotten settled in your work, what kind of living arrangements do you expect to have? (*Check one*)
 1 ______ I expect to live in the home of my parents (or my wife's parents)
 2 ______ I expect to live in the home of other relatives or friends
 3 ______ I expect to live in a rooming or boarding house
 4 ______ I expect to rent an apartment
 5 ______ I expect to rent a house
 6 ______ I expect to live in my own home
 7 ______ Other. *What?* ______________________________

67. After you have gotten settled in your work, how much rent or home payment do you think you will be able to pay per month? (*Check one*)

1 ______ Under $10.00 per month
2 ______ $10.00 to $14.99
3 ______ 15.00 to 19.99
4 ______ 20.00 to 24.99
5 ______ 25.00 to 29.99
6 ______ 30.00 to 39.99
7 ______ 40.00 to 49.99
8 ______ 50.00 to 59.99
9 ______ 60.00 or more
11 ______ I don't plan to pay rent or home payment

68. In this questionnaire we have asked you a lot of questions about your plans for after your discharge. Now that you have thought them over a little more, try to take everything into consideration and check the *one* thing you will most probably do *first of all*—after you have gotten your discharge and taken a vacation.

1 ______ I will probably work for the employer I was working for just before I came into the Army
2 ______ I will probably work for salary, wages, or commission for some other employer
3 ______ I will probably go to full-time school or college
4 ______ I will probably farm for myself or with my family
5 ______ I will probably go into business for myself
6 ______ I am undecided what I will do

69. How sure do you feel that this is what you will try to do *first of all?* (*Check one*)

1 ______ I am very sure
2 ______ I am pretty sure, but might do something else
3 ______ Chances are about 50–50 that this is what I will try to do
4 ______ I'm really not at all sure what I will do

70. About how long a vacation do you plan to take before you get down to doing whatever you expect to do? (*Check one*)

1 ______ I don't expect to take any vacation
2 ______ Less than two weeks
3 ______ Two weeks up to one month
4 ______ One month up to two months
5 ______ Two months up to three months
6 ______ Three months up to six months
7 ______ Six months up to a year
8 ______ A year or more

71. Do you plan to go back to live in the same place—the same local area, say within 5 or 10 miles of where you lived before you entered the Army? (*Check one*)

1 ______ Yes, I plan to live there for a long time to come
2 ______ Yes, I plan to go back there, but only for a short time
3 ______ No
4 ______ Undecided

72. What kind of a place do you plan to live in right after you leave the Army? (*Check one*)

1 ______ On a farm
2 ______ In the country but not on a farm
3 ______ In a village of less than 2,500 people
4 ______ In a small city (2,500 to 25,000 people)
5 ______ In a city of 25,000 to 100,000 people
6 ______ In a large city (more than 100,000 people)

73. How many points do you have under the Army Point System—your points for length of service, combat, children, etc., under the Score Card Plan?

I have ______________________ points

______ I don't know how many points I have

Appendix B. Questionnaire Used for Follow-up of July Separatees

(*Please read every question through before you make your answer. Mark some answer to every question. If any additional information is needed to explain your answer, please write it in.*)

(*You need not sign your name on this questionnaire unless you want to.*)

1. Which of these things are you doing at the present time? (*Check the answer or answers that apply to you—like this* √)

 1 ______ Working for salary, wages or commission
 2 ______ Working in your own business (not *being paid a salary by anyone else*)
 3 ______ Farming—either for yourself, with your family, or for wages
 4 ______ Going to full-time school
 5 ______ Just resting, visiting friends, etc.
 6 ______ Looking for a job
 7 ______ Something else. *What?* ______________________________

2. How many jobs have you had since you were discharged from the Army? (*Don't count it as a different job unless you left one company or employer and went to work for another company or employer. If you have done different kinds of work for the same company or employer without ever leaving, just count that as one job.*)

 1 ______ I have not taken a job yet
 2 ______ I have had only one job
 3 ______ I have had two jobs
 4 ______ I have had three or more jobs

3. If you are working, what kind of work are you doing now? (*Please tell what your job is called and just what you do.*)

 ______ I am not working

 ____

 ____

 ____

4. About how long were you out of the Army before you started working?

 __________ months and __________ days

 ______ I have not taken a job yet

5. Are you planning to keep the job you now have? (*Check one*)

 1 ______ I am not working
 2 ______ Yes, I plan to keep the job I now have, if I can
 3 ______ I will probably keep this job for a while, but may change jobs
 4 ______ No, I plan to leave the job I now have

 If you plan to leave the job you now have, what will you do?

 ____

 ____

6. Since you got your discharge, did you go back to work for the employer (company, person, etc.) you worked for before you entered the Army? (*Check one*)

 1 ______ I wasn't working for an employer before I went into the Army
 2 ______ Yes, I got a job with my old employer and I am working there now

3 ______ Yes, I got a job with my old employer but I am *not* working there any more
4 ______ I tried to get a job with my old employer, but couldn't
5 ______ No, I asked about a job but decided not to work there
6 ______ No, I haven't started working there yet, but I plan to go back with my old employer
7 ______ No, I didn't try to go back and I don't plan to

7. (a) Have you tried to get a job with any branch of the government (*Federal, state, city, etc.*) since you got out of the Army?
1 ______ Yes, and I am already working for a branch of the government
2 ______ Yes, I have my application in for a government job and will take the job if they offer it to me
3 ______ I did apply for a government job, but am not sure I want it now
4 ______ No, I haven't tried yet, but I plan to
5 ______ No, I am not interested in a government job

(b) If you have tried to get a government job, what job did you apply for and what did you do to try to get the job?

____
____
____
____

8. (a) Do you plan to go to *full-time* school or college? (*Check one*)
1 ______ I am going to *full-time* school or college *now*
2 ______ Yes, I plan to go to *full-time* school or college
3 ______ I may go to *full-time* school or college, but I'm not sure
4 ______ No, I don't plan to go to *full-time* school or college

(b) What school or college will you go to?

____

(c) When do you plan to start school or college?
Month ____________________ Year ____________________

9 If you are farming, whose land are you farming on? (*Check one*)
1 ______ I am not farming
2 ______ My family owns or rents the land and I am farming with them
3 ______ I own the land in my own name
4 ______ I am renting the land
5 ______ I am working on someone else's farm for wages

10. Have you made any arrangements for buying or starting a business of your own?
1 ______ No, I am not planning to have a business of my own
2 ______ I already have a business of my own
3 ______ I have been thinking of getting my own business, but have not made any arrangements yet
4 ______ Yes, I have made certain arrangements for getting a business of my own. *What arrangements?* ______________________________
____
____

11. If you were planning to have your own business, have you run into any difficulties you had not expected while you were in the Army?
1 ______ I wasn't planning to have my own business
2 ______ No, I have not run into any difficulties I hadn't expected
3 ______ Yes, I have run into some difficulties. *What difficulties?* __________
____
____

12. If you are still planning to have a business of your own, when do you plan to get started? (*Just a rough estimate*)

    Month ________________ Year ________________

    ______ I am not planning to have a business of my own

13. (a) Have you applied for any of the benefits of the GI Bill of Rights? (*Check one*)

    1 ______ Yes
    2 ______ No, but I plan to apply
    3 ______ No, and I don't plan to apply

    (b) If you have applied, check the benefits you applied for:

    1 ______ Loan to buy a house
    2 ______ Loan to go into business
    3 ______ Loan to buy a farm
    4 ______ Approval of aid for education or training
    5 ______ Unemployment payments

    (c) Did you receive the aid you applied for?

    1 ______ Yes
    2 ______ I haven't heard yet
    3 ______ No, my application was turned down

14. Before you were discharged from the Army, you probably had some plans about the kind of work you would do right after you got out of the Army. Have you been able to carry out the plans you had? (*Check one*)

    1 ______ Yes, I have carried out the plans I had
    2 ______ I haven't tried to carry out my plans yet but still expect to
    3 ______ No, I tried to carry out my plans but haven't been able to
    4 ______ I changed my plans before I really tried to carry them out
    5 ______ I really didn't have any definite plans

15. In what way have you not been able to carry out your plans, and why was this?

16. If you would care to make any suggestions to help the Army do a better job at the Separation Centers, or if you have any comments about problems you ran into that other soldiers should be told about, please write them here.

PLEASE CHECK BACK TO SEE THAT YOU DID NOT SKIP ANY QUESTIONS

APPENDIX

SOME NOTES ON SAMPLING AND QUESTIONNAIRE ADMINISTRATION BY THE RESEARCH BRANCH[1]

The following pages sketch briefly some of the problems involved in sample selection and questionnaire administration by the Research Branch in its operations during the war. To the procedures most commonly used and here described there were many exceptions necessitated by special types of inquiries. For example, the experimental studies as described in Volume III, though not involving large cross-section samples, required much more meticulous planning in handling the situations of questionnaire administration than was necessary in the average survey.

The Research Branch made no new contributions to sampling theory, other than the development of a few new formulas for calculating standard errors with certain types of data.[2] The problem of securing a sufficiently representative sample of the Army universe covered by a particular study was not an easy one, but its solution did not lead to new mathematical constructs. Because the situations in which sampling was done in the Army were quite different from any which would be encountered in civilian life, it has seemed unlikely that the Research Branch experience in sampling would prove especially instructive for future civilian research. Therefore, the description of the problems encountered and methods used is kept very brief here and is presented mainly as a service to those readers of preceding volumes who wish to know a little more about the kind of operations which lay behind the tables and charts presented there.

To many who worked in the Research Branch it soon became evident that error or bias attributable to sampling and to methods of questionnaire administration were relatively small as compared with other types of variation—especially variation attributable to different ways of wording questions.[3]

In the practical situation in which the Research Branch operated it was necessary to develop sampling procedures which were (1) swift, (2) economical, and

[1] By Samuel A. Stouffer. Acknowledgment is due to A. J. Jaffe for providing the principal data on sampling, to F. Douglas Williams for a memorandum on field work from which quotations are made, and to Shirley A. Star for data bearing on special methodological problems involved in studying Negro troops.

[2] For example, some simple formulas for testing significance in the case of order statistics due to Frederick Mosteller. See Volume I, Chapter 4, pp. 156 and 203. Another example is a formula for testing the significance of the difference between two correlated propositions, developed in the Experimental Section of the Research Branch. See Volume III, Appendix A; also Volume I, Chapter 4, p. 208.

[3] See, for example, Volume I, pp. 44–46.

(3) seemed to be within the limits of error tolerable for the particular study involved.

The sampling designs in the initial studies made by the Research Branch were developed mainly by Rensis Likert, Quinn McNemar, and S. A. Stouffer, the first committee called in by General Osborn to advise on research procedures. As the scope of studies increased, a permanent sampling section was set up, headed by A. J. Jaffe. When Jaffe went overseas he was succeeded by George Hausknecht and later by Robert Dubin, and upon his return from overseas resumed responsibility. Perhaps no one contributed more to the basic philosophy of the Research Branch sampling than Frederick Mosteller, who was consultant on sampling throughout the period when the Branch was shifting over from local planning surveys, exclusively, to cross-section surveys as well. Louis Guttman also helped in this field, as well as in scale construction, and was responsible for the sample design in several studies.

Although—perhaps one should say *because*—several of these men were expert statisticians, they kept their eye on practicalities. If it didn't matter too much whether a 5 per cent bias appeared in the data, there was no point in dragging out a study several weeks or doubling or trebling the man power involved on it in order to achieve greater sampling precision. In fact, as experience grew, there was increasing confidence in the feasibility of "cutting corners" in sampling with no serious loss to the ultimate reliability of the findings. On many topics the correlations between attitudes and those "demographic" characteristics which could be controlled by stratified sampling were so low that a sampling with large known bias in "demographic" characteristics would tend to yield about the same attitude responses as one with only a small bias. On other topics more caution was necessary. Consequently, the sampling design, with the accumulation of experience, was tailored more and more closely to the type of attitude data involved in a given study.

Ordinarily, sampling was a two-step process. First, there was sampling of organizations. This selection of sample points can be described as purposive or quota sampling. Second, within an organization there was sampling of individuals. The selection of individuals at a given sample point ideally involved procedures which might be called probability sampling.

In the earliest studies and in many of the later studies, the universe to be sampled consisted of a single organization, such as an Infantry division. That particular organization may have been selected for one of many reasons, which did not necessarily involve application of any explicit sampling principles. For example, a study was made in November 1943 of the First Infantry Division, selected because it was the only division with combat experience in the European theater at the time. Later, similar studies were made of three veteran combat divisions in the Pacific which were selected personally for study by the Chief of Staff of the Army. Obviously, generalizations from such studies to other divisions must be made with greater caution than if the data had been based on a random or stratified sampling of many units from all veteran combat divisions throughout the Army. But *within* a given division or organization the special Army situation offered facilities for probability sampling which often surpassed those in civilian life. If one wanted to get a representative sample of a particular organization, that was a relatively simple matter—in the ideal case.

Sampling of Individuals within a Unit

Let us discuss, first, the process of sampling individuals within a designated organization, and defer the discussion of selection of organizations.

Assume that the study design called for, say, 250 men in a particular organization.

The standard procedure was to draw the names of every nth man from a duty roster. The men whose names were drawn would then be ordered to appear at designated places, like mess halls, at designated hours, for administration of the questionnaire in groups of 25 to 50. In the ideal case, this situation was a sampler's dream.

Or, if the study design called for n_1 men with a high school or college education and n_2 without, or for n_1 with high AGCT scores and n_2 with low scores, or for certain numbers at a given rank or with a given age, the problem was only slightly more difficult. At the personnel office was a file of the soldiers' Form 20 cards containing all such information and much else. For any desired category of men these cards could be sampled at random and the men whose names were drawn could be ordered to appear at a designated time and place.

Thus, ideally, the military situation was almost perfect for sampling within units. But only in the ideal case. In practice, the ideal was never quite realized and frequently the local situation was such as to require improvisation on the part of the field team.

In order to appreciate the problem as it appeared in practice, one must understand the nature of the authority under which the Research Branch operated. General Osborn had full authority to make surveys, but he or officers under him could not give orders to commanders in the field. Whether a local commander could flatly refuse to cooperate was a problem which was never put to a definitive test. Almost always, when the Research officers approached a commanding officer at a post or with a unit, cooperation was forthcoming, sometimes grudgingly, but nevertheless it was given. Tact, dignity, and persuasiveness were often required on the part of the research officer who headed a team, since the commanding officer very often would suggest modifications of the sampling design, in the interest of minimum disruption of his troops' time schedule. In a few instances where commanding officers initially refused cooperation, the research officer, after all else failed, was able to secure cooperation by using words to this effect: "I am sorry, General X, but your units were selected in the Army cross-section sample and I have no authority to make substitutions. I can only call the War Department and tell them that you find it impossible to cooperate and await further instructions from Washington. Incidentally, this will be the first time in our experience that this has been necessary." Such a statement was usually enough, for the average commanding officer did not want to risk unknown trouble in Washington.

There were times, however, when the commanding officer had very compelling reasons either for not complying or for suggesting modifications in the sampling scheme or the proposed time schedule. For example, the organization might be out in the field on a tactical problem, scattered over a wide area in small units. To assemble a sample of men at the Research Branch's convenience might completely disrupt the training schedule. The Research Branch field team was instructed to make every reasonable effort to adapt its own time schedule to the local situation. Sometimes this meant returning to the camp a week later. Sometimes the time urgency in Washington was such that it was now or never. In this case, some discretion was permitted. If there was an a priori reason to think that a new unit could be substituted for the unavailable unit without bias, a substitution was made.

Obviously, the Research officer had to be tactful, resourceful, and conscientious as well as thoroughly aware of the hundred and one pitfalls which could produce bias in the sample he was sent to get. Here is a concrete example of a situation calling for a combination of all these qualities:[4]

[4] From a memorandum by F. Douglas Williams.

A Research Branch in an overseas theater agreed to undertake a survey—for a theater staff section—among casuals in a Replacement Depot awaiting shipment to the United States. It was further agreed that the sample would be a cross section of returnees representative of the theater, according to command, or a set of subsamples which could be correctly weighted to constitute such a cross section.

In talking with the adjutant of the Depot, the field officer was told that if every *n*th man were to be taken from a shipping list of men who had been processed, and were now awaiting ships, a thoroughly representative cross section of the Depot would be obtained—and, further, he knew men from all commands in the theater would be represented in approximately their proportionate strength.

This looked good, on paper. Investigation indicated, however, that the main interest of the adjutant in this system was that the survey would be restricted to men whose processing had been completed, and therefore the Depot procedure wouldn't be upset. Further investigation showed that the number and kind of men, and the command from which they came, present in the Depot at any given time varied considerably. Because of transportation procedure and scheduling of troops *to* the Depot, any shipping list might have been made up primarily of Ground Force men, or Air Force men, etc.

It thus became apparent that a sample from the whole depot (including men who had not been processed) would have to be drawn, if there were to be even a chance at all that the various commands in the theater would each be adequately represented.

The next step was for the Depot adjutant to suggest that he himself call each battalion commanding officer, and instruct him to order each of his company officers to select every *n*th man (the *n* of course to depend on the company strength) from a roster of the men who would be available at the time of the class. He said some men wouldn't be available because of being on KP, on guard, on pass, etc., but that all men received these things on the same rotational basis, in which there was no selective factor.

The field officer persuaded the adjutant that he could save the latter a lot of time and trouble if he (the field officer) visited the battalions himself, and made his own arrangements.

Before he was through, the field officer had to go to the companies in the battalions, in order to decide on the sampling method that could apply there. The headquarters of the battalions claimed to know, but didn't always, just what kind of a situation existed in their different companies. Even companies in the same battalion had different practices.

For instance, some kept up-to-date rosters, and some did not. This made a difference as to whether every *n*th man was selected from a paper record, or from an actual formation of the men.

In some companies all men drew details alike. In other companies sergeants drew no details, or only one or two particular kinds. In some companies noncoms received more passes than privates; in other companies this wasn't true Factors like this made a difference in whether the sample could be selected from the men with no other prearranged occupation (such as guard or KP) at the time of the questionnaire, or whether the sample would have to be drawn from all men in the company, in order that it be representative.

By this time another factor had become apparent. There were no adequate records in Depot, battalion, or company headquarters which could show the extent to which the various theater commands were represented. The Depot had a fluid, swiftly changing population, and their paper work wasn't designed to keep track of matters like this. There was no basis for stratifying even approximately by command, or stratifying to get a minimum number of men from certain commands.

Accordingly, the field officer decided to get a representative cross section of the Depot as of a particular time period. Then, at least, the study could be reported on that basis. He also increased the size of the total sample, with the hope that then even the commands with the smallest number of cases would be large enough to permit comparing them with the commands having larger representation. (Weighting could of course be used to arrive at theater averages.) Happily, his hope was borne out in practice.

Fortunately, most situations were not as complicated as this, but seldom was a survey conducted which did not involve problems. Inexperienced Research officers learned rapidly after happenings like the following:[5]

A staff officer in an organization overseas who was very cooperative was permitted, contrary to Research Branch instructions, to select the *n*th men himself from the rosters. When he started to check the names he found that one third of the unit's strength was away at a rest camp. So he just substituted names at will. The research officer learned of this after the men had filled out the questionnaires. He wasn't in a position to find out much about the basis on which the staff officer had substituted names, and, further, the sample couldn't have been good with that many men absent (selective factors were present in determining which men from the unit had been given furloughs). All the Research officer could do was discard the questionnaires and substitute another unit.

In another case, the carrying out of a sampling procedure consisted of a battalion commanding officer instructing his adjutant to inform all the company commanders to check every *n*th name off a roster, so that such and such a number of men could be selected. The Research Branch officer was too far removed from this operation to control it, and he later found out that in practice the prescribed number of men were just arbitrarily picked from some of the companies and told to report for the questionnaire. Again, the questionnaires from that battalion had to be discarded, and another organization substituted for it.

In the third case, the field officer wanted the total strength of a unit for his sample. The questionnaire was a short one, and the executive officer convinced the research officer that a special order or announcement for a particular time for the men to take the questionnaire would not be necessary. Rather, it could be administered during the Orientation Hour, which all men always attended. Somewhat more than a half of the men turned out to be actually present, and accordingly a biased and insufficient sample resulted. The completed questionnaires were of course useless from this unit, and they were eventually discarded and a new unit substituted.

An illustration of the way in which quotas at a particular installation were assigned and the extent to which these quotas were filled is given in Table 1. Camp Stewart, Georgia, was one of seventeen installations visited in February 1944, in connection with Survey 95. The specific units at Camp Stewart were preselected in Washington and quotas assigned as shown in the first column of Table 1, indicating the number of men to be called. Except in the case of the 1st Provisional Cadre, the men were to be selected from rosters by the *n*th man procedure; allowance was made for about 20 per cent absentees, due to leaves, hospitalization, etc., as shown by the expected numbers in the second column.[6] (In the case of the 1st Provisional Cadre, since it was desired to get overseas returnees from specified places, special lists of available men were to be drawn up at the camp and these lists sampled randomly.) As often happened, the information in Washington proved to be out of date. The 561st Anti-Aircraft Battalion had shipped out and the Research Branch officer exercised his discretion in substituting the 580th. However, the 82nd Finance Disbursing Section also had shipped out and no substitute unit was available at Camp Stewart. The Research officer promptly notified Washington, which added a comparable unit to the quota of another camp.

It will be seen from the third column that the number actually reporting was a little below estimates in the anti-aircraft battalions. In the case of the Service

[5] From Williams, *loc. cit.*

[6] Except in special instances, there was no reason to think that a normal percentage of absentees on leave introduced a serious sampling bias. Omission of those in the guardhouse did represent bias, of course, but such men were few in number. Omission of those hospitalized could introduce bias, especially if the complaint was psychoneurotic, but these cases of absentees from the roster also were relatively few.

Command units the number reporting was above expectation, partly because the Research Branch officer interpreted his instructions as permitting calling somewhat more than 45 or 80 names originally. Table 1 is used illustratively because it shows quite typical situations—neither better nor worse than what usually was found.

There were frequent studies requiring special samples which involved drawing the respondents from the Form 20 cards, which contained background information on the men. For example, a study might require an extra sample of college men only or of men with particular job skills or training. It was then necessary to sort the Form 20 cards and draw the sample randomly from subsets of these cards in-

TABLE 1

ASSIGNED QUOTAS AND EXTENT OF FULFILLMENT, CAMP STEWART, GA. (February 1944)

	Number to be called	*Average number expected to report*	*Number actually reporting*
Anti-Aircraft Training Command			
297th Anti-Aircraft Battalion	93	74	67
298th Anti-Aircraft Battalion	89	71	59
*561st Anti-Aircraft Battalion	95	76	69
589th Anti-Aircraft Battalion	111	89	75
590th Anti-Aircraft Battalion	97	68	63
Service Command Units			
1474 MP Detachment	About 45	36	45
1474 Supply Detachment	About 80	54	80
82nd Finance Disbursing Sect.	55	44	0**
Overseas Returnees			
910th Anti-Aircraft Battalion	70	56	68
1st Provisional Cadre			
Men from Caribbean	25	25	27
Men from Iceland-Greenland	25	25	25
Men from Africa-Sicily-Italy	25	25	20

* Shipped out 580th Anti-Aircraft Battalion substituted
** Shipped out. No comparable unit present. Hence no cases obtained

stead of from rosters. Wherever a unit, as was almost invariably the case with Negro units, was expected to contain a substantial number of illiterates or semiliterates, it was necessary to use the Form 20 cards instead of rosters. Thus the men to be personally interviewed could be segregated in advance and sent to special points where trained interviewers were awaiting them. Problems involved in personal interviewing will be discussed later in this Appendix.

Selection of Organizations To Be Included in a Study

In spite of all the difficulties encountered in the field, the Army situation usually made possible the obtaining of a satisfactory random sample of individuals within the units preselected. The greatest chance for bias was in the initial selection of the units.

As has been indicated, the preselection was done in Washington for troops in

the United States or at theater headquarters for troops in a given theater overseas. The Research Branch had access to the latest secret data on troop strength and location of units, although in some rapidly moving periods the information was out of date by the time the study went into the field.

The selection of units involved purposive rather than strictly random sampling. If a cross section of enlisted men in the United States or a theater was required, the standard procedure was to select the units such that each branch of service was represented in as nearly as possible the correct proportions. Within a given branch of service, effort was made to get units in various stages of training or with various types of Army experience in their correct proportions, in so far as this could be inferred from available data. For example, among infantrymen in the United States, some would be in Replacement Training Centers, some in newly activated divisions, some in divisions scheduled for early shipment overseas. The proportions of men in each type of unit could be roughly calculated, and the sample was designed to maintain these proportions.

Having stratified the universe by branch and type of unit, one could proceed to draw one or more units from each subcell of this stratified universe. Ideally, this drawing would be at random, but practical considerations usually prevented a strictly random selection of units within the subcell. Most important was the fact that the number of field representatives of the Research Branch, either at home or overseas, was limited. Therefore, it was economical and, in fact, necessary, to limit severely the number of posts to be visited. It was much faster to get a variety of units at each of a few posts than one unit each at a large number of posts. The number of posts visited depended on the character of the study. If the study called for data which would be highly correlated with post location or with particular post policies, many more posts would be needed than if the study called for data on, say, attitudes toward our allies or desire for overseas service. Usually, not more than a dozen or twenty posts were visited on a particular cross-section survey in the United States, and several different kinds of units might be sampled at each post.

The need for speed and economy led to a systematic bias in favor of selecting the larger posts rather than installations with only a few men. Statistical analyses of attitude data showed that there were no differences in most attitudes as between men in the two kinds of installations, except on occasional topics like leisure-time activities. If a survey emphasized the latter, for example, special effort was made to get an adequate sample of smaller installations.

Another practical consideration was the necessity of routing the field teams so that a given team could cover a maximum number of installations in a minimum time. Thus one team would go to Texas, Louisiana, and Mississippi; another to Alabama, Georgia, and Florida; another to Oregon and California; another to the Carolinas and Virginia; another to Indiana, Kentucky, and Tennessee; another to New England; etc. Railroad and air schedules actually were consulted before the final decision on camps was made.

Finally, since sometimes several surveys had to be made in quick succession, explicit effort was made to avoid bothering any one installation or any one unit within that installation any oftener than necessary.

When all these considerations were taken into account, it can be seen that a large element of purposiveness rather than randomness entered into the selection of installations to be visited.

Checking the Adequacy of the Sample after It Was Taken

After each survey, the sampling experts checked the returns to see whether large biases had occurred on variables for which data were available for the uni-

TABLE 2

COMPARISONS OF DEMOGRAPHIC CHARACTERISTICS IN RESEARCH BRANCH SAMPLE AND IN AGO 2 PER CENT SAMPLE OF ARMY
(Percentage Distribution)

	UNITED STATES		OVERSEAS		TOTAL	
	Research Branch	*AGO*	*Research Branch*	*AGO*	*Research Branch*	*AGO*
Rank						
Sergeant	28	26	30	28	29	27
Corporal	18	16	22	23	21	21
Pfc	21	21	36	34	31	30
Private	33	37	12	15	19	22
	100	100	100	100	100	100
Marital Condition						
Single	49	52	61	62	58	60
Married, no children	25	25	20	21	22	22
Married, 1 child	16	14	13	12	13	12
Married, 2 children	7	6	4	3	5	4
3 or more children	3	3	2	2	2	2
	100	100	100	100	100	100
Length of Time in Army						
Over 3 years	27	22	29	23	28	22
2 years up to 3 years	35	37	46	48	43	45
1 year up to 2 years	14	18	18	22	17	21
Under 1 year	24	23	7	7	12	12
	100	100	100	100	100	100
Length of Time Overseas						
Over 3 years	2	1	2	1	2	1
2 years up to 3 years	7	7	13	13	11	10
1 year up to 2 years	8	8	35	39	25	29
6 months up to 1 year	4	5	24	24	18	19
Under 6 months	1	3	26	23	18	17
Never overseas	78	76	—	—	26	24
	100	100	100	100	100	100
Age						
35 and over	12	13	8	10	9	11
30–34	17	17	16	17	16	17
25–29	31	31	31	33	31	33
20–24	26	29	39	38	35	35
Under 20	14	10	6	2	9	4
	100	100	100	100	100	100
Number of cases in Research Branch sample	*2,417*		*4,886*		*7,303*	

Research Branch sample, S-205, May 1945. AGO sample, comprising about 2 per cent of entire Army, March, 1945.

verse—such as rank, length of time in the Army, age, etc. In a few instances, the sample revealed a bias—not including quite enough men in the early stages of training, for example, or including too many. When feasible, an additional sample was drawn immediately to correct such a bias.

In checking the reliability of its samples, the Research Branch was severely hampered, especially in the early years of the war, by the absence of up-to-date demographic data on the universe. From the beginning, the Branch had used all the influence it could muster to get the Army to adopt a more flexible bookkeeping system as to its personnel. Experts from the Branch were loaned to other Army agencies at various times to help achieve this. Especially helpful was a periodic 2 per cent sampling of the entire Army instituted by the Machine Records Branch of the Adjutant General toward the end of the war. While this sample was never entirely complete, and had some biases which the Research Branch statisticians helped discover, it was invaluable in providing a check on some of the important demographic characteristics of Research Branch data such as length of time in the Army and length of time overseas.

An example of the extent to which demographic characteristics as derived from a 2 per cent AGO sample taken as of March 1945 and a Research Branch cross section (S-205) as of May 1945 agreed, is shown in Table 2. In most characteristics, as Table 2 shows, the agreement is quite close, though some discrepancies are statistically significant. If we assume (as is not entirely justified) that the AGO sample is correct, there is probably no attitude item in which overall proportions of respondents in any *attitude* category would be changed by more than 1 or 2 per cent by reweighting attitude responses according to proportions in a given AGO sample category rather than according to proportions in the Research Branch sample.

TABLE 3

SCORES UNDER THE POINT SYSTEM AS REPORTED BY MEN IN RESEARCH BRANCH SURVEY AND AS ESTIMATED BY AGO FROM OFFICIAL RECORDS, MAY 1945
(Percentage Distribution)

	UNITED STATES		OVERSEAS		TOTAL	
	Research Branch	*AGO*	*Research Branch*	*AGO*	*Research Branch*	*AGO*
85 and over	9	8	14	11	13	10
75–84	6	4	11	9	9	7
65–74	5	5	16	13	12	10
55–64	4	6	16	18	13	14
45–54	9	9	16	19	13	16
Under 45	67	68	27	30	40	43
	100	100	100	100	100	100

Research Branch data from S-205.

Table 3 compares the distribution of point scores as reported by the men in the same Research Branch survey with a distribution of point scores as determined independently by the AGO from official records.

Surveys throughout the war were checked routinely against available Army data to determine the extent of bias in the sampling. In view of the magnitude of the variations in response which could be induced by even small changes in the

wording of questions, the errors introduced by sampling were generally thought to be so small as to be relatively unimportant.

Methods of Questionnaire Administration

Another potential source of error was in the process of administration of the questionnaires. Normally, a sample of about fifty enlisted men were assembled by order of their unit commander at a place convenient for writing—a mess hall, for example. The class leader was almost always an enlisted man, trained for the purpose. His opening remarks would go somewhat as follows:

You are going to be given some questions to answer. You can answer them just exactly the way you personally feel about them. There are no right or wrong answers.

Believe it or not, the Army is interested in knowing how you personally feel about some things in the Army. And in order to be sure that you feel free to say what you think just as frankly as you want to say it, these questionnaires are anonymous. Do not sign your name or serial number.

When you have finished, put your questionnaire in a pile on the table. Your CO will not see any of these questionnaires.

You may have been wondering why you were selected for this. Well, out of the thousands of units in the Army a few units are pulled out of the hat, as it were, from the Army. What men in these outfits say will correspond very closely with what all the men in the Army will say. And we also have drawn men at random from the roster in your outfit. So your name came up by chance and here you are.

These questionnaires will be sent to Washington (a theater headquarters if overseas) and tabulated in statistical machines. Nobody will ever be able to identify your questionnaire. Here's your chance to say what you really think!

Qualitative evidence of validity of the men's responses can be derived from inspection of the filled-out questionnaires. With few exceptions, the hundreds of thousands of men throughout the world who filled out questionnaires took the task with obvious earnestness. There can be little doubt about their frankness. And the fact that responses both "hung together" internally and were correlated with outside criteria has been demonstrated in the many tables and charts presented in Volumes I and II.

This is not to say that men were not skeptical as to the extent which the Army would pay attention to what they wrote. Some of them said quite frankly that they didn't expect the Army to do anything about the situations covered in the questions; but such cynicism did not inhibit them from answering, and, indeed, from elaborating their remarks in very detailed free comments written on the margins of the questionnaires as well as in spaces provided for such comments.

Several studies were conducted to test explicitly the variations in response introduced by various methods of questionnaire administration. Since the Army situation has few parallels with that in civilian interviewing, which ordinarily cannot take advantage of the economies of "classroom" administration, these studies will be reviewed only in brief summary.

Before the first planning survey was made in December 1941, a methodological study was conducted to compare classroom administration and personal inter-

[7] This was rigorously adhered to. An incident which became something of a classic story in the Research Branch occurred early in the war. The commanding general of a post walked into a hall near the end of a session and started leafing through the questionnaires which the early finishers had already tossed on the table. The corporal in charge of the table said, "Sir, my orders are that no officer in this camp will be permitted to see these questionnaires." The general immediately apologized and, returning to headquarters, commended the officer in charge of the Research Branch team for having an enlisted man with such a sense of duty.

viewing. This study, made in November 1941, was designed as follows:

Group A filled out the questionnaire in the classroom.

Group B was interviewed personally and individually by trained enlisted men, using the same questionnaire.

Group C filled out the first half of the questionnaire in the classroom and was personally interviewed on the second half.

Group D was interviewed on the first half of the questionnaire and filled out the second half in the classroom.

Each group comprised fifty cases. No significant differences were found between any of the groups. Moreover, the respective halves of split questionnaires contained questions with parallel forms and the differences between responses to parallel forms by the same persons were not significant. The only apparent difference between classroom administration and interviewing results was the fact

TABLE 4

RELATION OF AGCT, EDUCATION, AND RACE TO NONRESPONSE ON SURVEYS, CLASS-ADMINISTERED QUESTIONNAIRES COMPARED WITH PERSONAL INTERVIEWS (February 1943)

DESCRIPTION OF GROUP			NUMBER OF MEN		AVERAGE PERCENTAGE OF UNANSWERED QUESTIONS*	
Race	*AGCT Level*	*Education*	*Classes*	*Interviews*	*Classes*	*Interviews*
White	I–IV	—	217	—	2.0%	—
Negro	I–III	—	96	—	1.0	—
Negro	IV	—	82	—	2.0	—
Negro	V	8th grade or more	47	74	2.5	1.5%
Negro	V	4th–7th grade	94	98	2.9	1.4
Negro	V	Under 4th grade	86	73	8.6	2 1

* This figure is the total number of nonresponses divided by the product of the number of men surveyed and the number of questions asked There were 59 background and attitude items common to the white and Negro questionnaires. The Negro questionnaire contained an additional 15 attitude items as well.

that more detailed free comments were written down by the men themselves than by their interviewers.

In this part of the study men with Class V (lowest) AGCT scores were not included; it was thought that they would probably have to be interviewed personally in any case. However, a group of such men were assembled in a classroom and questions were read to them aloud, giving each man time to check the answers on a questionnaire in front of him. The results checked closely with those obtained by individual personal interview, but the oral classroom procedure was so cumbersome and time-consuming that it was never used in practice.

In the first year of operations, all Class V men were interviewed routinely. Subsequent methodological studies showed that it was not necessary to interview the majority of such men as long as questions were simple in form. However, on complicated questions the results were better from personal interviewing. As time went on, and especially as pressure increased to get results quickly, the Research Branch tended to dispense with interviewing Class V men, except for a

few who when writing in the classroom had obvious difficulties in reading the questionnaire.[8]

Negro soldiers, as was reported in Volume I, Chapter 10, constituted a special problem. So many of them not only were in AGCT Class V, but also had had a most meager education, that any study of Negroes necessarily involved a substantial amount of personal interviewing.

Two quite elaborate methodological studies of the problem of eliciting information from Negroes were made, in January 1943 and in February 1943. Using as the criterion the number of questions actually answered, it was found that Negroes in all but AGCT Class V with less than 4th grade education handled the classroom questionnaires satisfactorily. This is shown in Table 4, which compares whites in AGCT Classes I to IV with Negroes in various subgroups. In the Negro

TABLE 5

RESPONSES BY NEGRO ENLISTED MEN FROM AGCT CLASS IV IN INTERVIEWS BY NEGROES AS COMPARED WITH CLASSROOM SITUATION AND IN INTERVIEWS BY NEGROES AS COMPARED WITH INTERVIEWS BY WHITES

	Elicited by Negro interviewers as compared with classroom situation	*Elicited by Negro interviewers as compared with white interviewers*
Excess in percentage of responses indicating racial protest (average of 4 questions)	+1%	+21%
Excess in percentage of responses indicating low personal commitment (average of 4 questions)	+2	+14
Excess in percentage of responses indicating lack of enthusiasm for war aims (average of 8 questions)	−6	+8
Excess in percentage of responses indicating pessimism about postwar conditions (average of 9 questions)	−3	+21
Excess in percentage of responses indicating unfair treatment in the Army (average of 5 questions)	+6	+16
Excess in percentage of responses indicating lack of high regard for officers and noncoms (average of 3 questions)	−4	+2
Excess in percentage of responses indicating relatively low personal esprit or job satisfaction in the Army (average of 13 questions)	0	+8

[8] Since practically all soldiers could speak and read English, the foreign language problem was negligible. In one instance, in Panama, an entire questionnaire was prepared in Spanish for Puerto Rican troops.

sample, where interviews were used, there were Negro enlisted men trained for the purpose.

Special attention was given to differences which might result in using Negro or white enlisted men as interviewers. A sample of 130 Negroes from AGCT Class IV, for example, was interviewed by Negroes; a sample of 146 Negroes matched with the 130 on AGCT, education, and other variables, was interviewed by whites. A third sample of 140 Negroes was given the same questionnaire in a classroom situation with Negro class leaders.

Table 5 shows that there was no essential difference in response between the results using Negro interviewers or classroom administration. But the differences in response to Negro and white interviewers, respectively, are quite consistently large. Negro interviewers tended more than white interviewers to elicit responses

TABLE 6

PROPORTION WRITING FREE COMMENTS ABOUT THE BUS SERVICE, CLASSIFIED BY OPINION OF BUS SERVICE, EDUCATIONAL LEVEL, AND RACE

RACE AND EDUCATIONAL LEVEL	PROPORTION VOLUNTEERING COMMENTS AMONG THOSE WHO SAID BUS SERVICE WAS:			
	Very good	*Fairly good*	*Poor*	*Very poor*
White				
Grade school	12 (*319*)	16 (*674*)	27 (*247*)	39 (*256*)
Some high school	21 (*132*)	26 (*481*)	41 (*240*)	46 (*266*)
High school graduate	22 (*151*)	31 (*490*)	48 (*296*)	56 (*277*)
College	35 (*91*)	39 (*226*)	51 (*125*)	60 (*133*)
Negro				
Grade school	22 (*660*)	26 (*890*)	46 (*842*)	54 (*1,299*)
Some high school	20 (*116*)	37 (*286*)	51 (*314*)	58 (*734*)
High school graduate	24 (*50*)	44 (*190*)	62 (*209*)	70 (*484*)
College	44 (*27*)	41 (*103*)	62 (*131*)	69 (*394*)

S-32, March 1943 Cross sections of white and Negro enlisted men in the United States. Number of cases in parentheses.

reflecting racial protest, unfavorable attitudes toward the war, pessimistic views of postwar conditions, unfavorable reports on the Army, and manifestations of low personal esprit. On the basis of such findings, the Research Branch routinely used Negro interviewers where interviewing was necessary.[9]

Who Wrote the Most Free Comments?

Studies made on many occasions brought out the fact that among all soldiers free comments tended to be made most frequently by (a) the better educated and (b) the most aggrieved with respect to a particular problem. This fact has been stressed frequently in Volumes I and II by way of counseling caution in the interpretation of free comments. Men were encouraged to write free comments, and, as already has been noted, the classroom situation was somewhat more productive

[9] It is not claimed that the response given to whites is an "untrue" response and that given to other Negroes is a "true" response, even if the latter did not deviate far from that obtained in the anonymity of the classroom. Both kinds of responses could be relevant, one being a "public" response to a white man, and the other a "private" response. Since in general only one type could be obtained in an actual cross-section study, the Research Branch decided to use Negro interviewers and Negro class leaders.

of free comments than the personal interview situation. But the lesser educated are likely in either case to be less articulate than the better educated. How education and sense of grievance combine to elicit free comments is shown in Table 6. Among white soldiers who checked "very poor" in response to a question about the bus service, 60 per cent of the college men wrote comments as compared with 39 per cent of the grade school men. Among those who checked "very good," 35 per cent of the college men wrote comments as compared with 12 per cent of the grade school men. Negroes, who had additional grievances against the bus service because it often was in fact poorer than for whites and because Jim Crow regulations were observed, were more likely than the whites to offer comments. This was true in 15 out of the 16 comparable categories in Table 6, by education and by opinion of the service.

The files of the Research Branch in Washington contain much data which further illustrate or document points mentioned in this Appendix. These records may be consulted by accredited social scientists who seek further information on the methodological problems discussed here.

Pressed as it was by the necessity of getting reasonably reliable information as quickly as possible, the Research Branch made many methodological compromises. Its sampling methods were by no means models of what sampling methods should be—although the Army situation was in some respects more favorable to the use of modern probability sampling methods (once the sample points were located) than is often possible in civilian life. Even more methodological studies would have been made to experiment with alternative interviewing situations, if time had permitted and personnel been available. As has been noted, however, the statisticians in the Branch were convinced that the biases introduced by sampling designs and their field execution were very small as compared with those attributable to the wording of questions. Hence, the concern of the Branch became more and more to improve its methods of question design—which led to the work on development of scaling theory occupying the first eleven chapters of the present volume, and to the analysis of outcomes of prediction using different types of items, as is described in Chapters 12 to 16.

Progress in theory and research with respect to sampling and interviewing has been rapid since the war. In this work government agencies like the Bureau of the Census have played a leading role, along with academic and business research organizations. Forthcoming monographs prepared under the auspices of the Joint Committee of the National Research Council and the Social Science Research Council, on the Measurement of Opinions, Attitudes, and Consumer Wants, may be expected to add further to our knowledge.

INDEX

A

B

C

D

E

F

G

H

I

J

K

L

N

O

P

S

T

U

XYZ

THE SOCIAL SCIENCE RESEARCH COUNCIL was organized in 1923 and formally incorporated in 1924, composed of representatives chosen from the seven constituent societies and from time to time from related disciplines such as law, geography, psychiatry, medicine, and others. It is the purpose of the Council to plan, foster, promote, and develop research in the social field.

CONSTITUENT ORGANIZATIONS

American Anthropological Association
American Economic Association
American Historical Association
American Political Science Association
American Psychological Association
American Sociological Society
American Statistical Association